Probabilistic Structured Models for Plant Trait Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Farideh Fazayeli

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ARINDAM BANERJEE

March, 2017

© Farideh Fazayeli 2017
ALL RIGHTS RESERVED

Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Arindam Banerjee, for his continuous support of my PhD study and research. He has been a tremendous researcher, mentor, instructor, and role model. His patience, motivation, enthusiasm, immense knowledge, and dedication to research encourage me to never give up and overcome difficult obstacles, and trigger me to grow as a research scientist. I have been extremely lucky to have him as my PhD advisor and will be forever thankful to him for introducing me to the wonders of scientific research.

I would like to thank Prof. Sudipto Banerjee, Prof. Daniel Boley, Prof. Vipin Kumar, and Prof. Peter B. Reich for being my dissertation committee members. I am also grateful to my collaborators Jens Kattge, Peter B. Reich, Franziska Schrodt, Habacuc Flores-Moreno, Abhirup Datta, Anuj Karpatne, Ethan Butler, Kirk Whythers, and Ming Chen. It was a pleasure to work and collaborate with them on the plant trait analysis throughout almost four years of the project.

I will miss my long discussion sessions and interactions with all my labmates and friends: Amir Taheri, Andre Goncalves, Hardik Goel, Huahua Wang, Igor Melnyk, Jamal Golmohammady, Karthik Subbian, Konstantina Christakopoulou, Miao Fan, Nicholas Johnson, Puja Das, Qilong Gu, Robert Giaquinto, Shaozhe Tao, Sheng Chen, Sijie He, Soumyadeep Chatterjee, Vidyashankar Sivakumar, Xiaoli Liu, Yingxue Zhou.

My special appreciation goes to my family for their endless love in my whole life. It would have been impossible for me to finish this work without their encouragement, understanding, support, and help.

I owe my loving thanks to my husband and best friend, Hamed Kajbaf, for all the emotional support, camaraderie, enthusiasm, caring, and helping me to get through the difficult times.

Dedication

To my parents and Hamed.

Abstract

Many fields in modern science and engineering such as ecology, computational biology, astronomy, signal processing, climate science, brain imaging, natural language processing, and many more involve collecting data sets in which the dimensionality of the data p exceeds the sample size n. Since it is usually impossible to obtain consistent procedures unless $p < n$, a line of recent work has studied models with various types of low-dimensional structure, including sparse vectors, sparse structured graphical models, low-rank matrices, and combinations thereof. In such settings, a general approach to estimation is to solve a regularized optimization problem, which combines a loss function measuring how well the model fits the data with some regularization function that encourages the assumed structure.

Of particular interest are structure learning of graphical models in high dimensional setting. The majority of statistical analysis of graphical model estimations assume that all the data are fully observed and the data points are sampled from the same distribution and provide the sample complexity and convergence rate by considering only one graphical structure for all the observations. In this thesis, we extend the above results to estimate the structure of graphical models where the data is partially observed or the data is sampled from multiple distributions. First, we consider the problem of estimating change in the dependency structure of two p-dimensional models, based on samples drawn from two graphical models. The change is assumed to be structured, e.g., sparse, block sparse, node-perturbed sparse, etc., such that it can be characterized by a suitable (atomic) norm. We present and analyze a norm-regularized estimator for directly estimating the change in structure, without having to estimate the structures of the individual graphical models. Next, we consider the problem of estimating sparse structure of Gaussian copula distributions (corresponding to *non-paranormal* distributions) using samples with missing values. We prove that our proposed estimators consistently estimate the non-paranormal correlation matrix where the convergence rate depends on the probability of missing values.

In the second part of thesis, we consider matrix completion problem. Low-rank

matrix completion methods have been successful in a variety of settings such as recommendation systems. However, most of the existing matrix completion methods only provide a point estimate of missing entries, and do not characterize uncertainties of the predictions. First, we illustrate that the the posterior distribution in latent factor models, such as probabilistic matrix factorization, when marginalized over one latent factor has the Matrix Generalized Inverse Gaussian ($\mathcal{MGIG}$) distribution. We show that the $\mathcal{MGIG}$ is unimodal, and the mode can be obtained by solving an Algebraic Riccati Equation equation. The characterization leads to a novel Collapsed Monte Carlo inference algorithm for such latent factor models. Next, we propose a Bayesian hierarchical probabilistic matrix factorization (BHPMF) model to 1) incorporate hierarchical side information, and 2) provide uncertainty quantified predictions. The former yields significant performance improvements in the problem of plant trait prediction, a key problem in ecology, by leveraging the taxonomic hierarchy in the plant kingdom. The latter is helpful in identifying predictions of low confidence which can in turn be used to guide field work for data collection efforts.

Finally, we consider applications of probabilistic structured models to plant trait analysis. We apply BHPMF model to fill the gaps in TRY database. The BHPMF model is the-state-of-the-art model for plant trait prediction and is getting increasing visibility and usage in the plant trait analysis. We have submitted a R package for BHPMF to CRAN. Next, we apply the Gaussian graphical model structure estimators to obtain the trait-trait interactions. We study the trait-trait interactions structure at different climate zones and among different plant growth forms and uncover the dependence of traits on climate and on vegetation.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables ix

List of Figures xii

1 Introduction 1

1.1 Structured Graphical Models 2

1.1.1 Structure Learning of Graphical Models 3

1.1.2 Low-rank Matrix Completion 4

1.2 Plant Trait Analysis 5

1.3 Overview and Contributions 7

1.3.1 Generalized Direct Change Estimation 8

1.3.2 Gaussian Copula Precision Estimation with Missing Values . . . 10

1.3.3 Collapsed Monte Carlo Inference for Matrix Completion 11

1.3.4 Matrix Completion with Hierarchical Side Information 12

1.3.5 Trait-Trait Interactions across Climate Zones 13

2 Related Work 15

2.1 Structured Learning of Graphical Models 15

2.1.1 Gaussian Graphical Models 15

2.1.2 Ising Graphical Models 18
2.1.3 Direct Change Estimation 20
2.2 Low Rank Matrix Completion 20
2.2.1 PMF, PPCA, and Bayesian PCA 21

I Structure Learning of Graphical Models **24**

3 Generalized Direct Change Estimation in Graphical Models **25**
3.1 Introduction 25
3.2 Generalized Direct Change Estimation 26
3.2.1 Ising Model 26
3.2.2 Loss Function 27
3.2.3 Optimization 29
3.2.4 Regularization Function 30
3.3 Theoretical Analysis 32
3.3.1 Background and Assumption 32
3.3.2 Bounds on the regularization parameter 34
3.3.3 RSC Condition 35
3.3.4 Statistical Recovery 36
3.4 Experiments 37

4 Gaussian Copula Precision Estimation with Missing Values **40**
4.1 Introduction 40
4.2 Method 41
4.2.1 Kendall's tau with missing values 42
4.2.2 Spearman's rho with missing values 43
4.2.3 Plugin estimate for CLIME 43
4.3 Theoretical Analysis 44
4.3.1 Kendall's Tau with Missing Values 46
4.3.2 Spearman's Rho with Missing Values 47
4.3.3 Plug-in CLIME Estimator 51
4.4 Experimental Results 52

4.4.1 Synthetic Data . . . 53
4.4.2 Climate Data . . . 55

II Low Rank Matrix Completion **57**

5 Collapsed Monte Carlo Inference for Matrix Completion **58**
5.1 Introduction . . . 58
5.2 Background and Preliminary . . . 59
5.2.1 Importance Sampling . . . 60
5.2.2 MGIG Distribution . . . 60
5.2.3 Algebraic Riccati Equation . . . 63
5.3 MGIG Properties and Sampling . . . 65
5.4 Connection of MGIG and Bayesian PCA . . . 67
5.4.1 Closed form Posterior Distribution in Bayesian PCA . . . 67
5.4.2 Posterior Distribution with Missing Data . . . 68
5.4.3 Collapsed Monte Carlo Inference for PMF . . . 69
5.5 Experimental Results . . . 71
5.5.1 Datasets . . . 71
5.5.2 Methodology . . . 72
5.5.3 Results . . . 72

6 Matrix Completion with Hierarchical Side Information **77**
6.1 Introduction . . . 77
6.2 BHPMF . . . 78
6.2.1 Model specification . . . 79
6.2.2 Sampling U . . . 79
6.2.3 Sampling V . . . 80
6.2.4 BHPMF Inference . . . 81
6.3 Experimental Results . . . 82
6.3.1 Dataset . . . 82
6.3.2 Baselines . . . 83
6.3.3 Methodology . . . 84

6.3.4 Results 85
6.4 Multiple Inheritance BHPMF 90

III Application **93**

7 Trait-Trait Interactions across Climate Zones **94**
7.1 Statement of Contribution of co-authors 94
7.2 Introduction 94
7.3 Method 98
7.3.1 Data 98
7.3.2 Analysis 100
7.3.3 Results 103
7.4 Discussion 105
7.4.1 Connectivity across all terrestrial plants 106
7.4.2 Trait connections across growth forms and climate regions . . . 107
7.4.3 Modularity 108
7.4.4 Modularity across climate regions and growth forms 108
7.4.5 Modularity across a precipitation and a temperature gradient . . 109
7.4.6 Plant trait network analyses on a precipitation and temperature gradient holding the other environmental variable constant across each gradient 111
7.4.7 Calculation of latent variables using a sparse precision matrix and low rank matrix 112

8 Conclusion **126**

References **129**

Appendix A. Direct Change Estimation Appendix **152**
A.1 Background and Preliminaries 152
A.1.1 Generic Chaining 154
A.2 Regularization Parameter 156
A.3 RSC condition 166

List of Tables

4.1 Edges dicovered by DoPinG and mGlasso on Climate Data. > denotes the number of edges in DoPinG graph but not in mGlasso graph. < is on the contrary. 56

5.1 Time Comparison of CMC and MCMC on different datasets. At each step of MCMC, rows of U and V can be sampled in parallel denoted by MCMC parallel. The running time is reported over 1000 steps for both methods where MCMC has 200 effective samples and CMC has 1000 effective samples. Note that the effective number of samples of MCMC is less than 1000 and more steps is required to obtain enough samples. The number of iterations for convergence of CMC is much less than 1000 (Figure 5.5). 76

6.1 ID, name, percentage of missing entries (%) and definition of the respective trait. 84

6.2 RMSE of Species Mean, PMF, HPMF and BHPMF. Latent dimension k=15 for matrix factorization methods. 85

7.1 Examples of studies focused on multi-organ, multi-trait datasets. When several plant group classifications were used, a semicolon divides them. The numbers next to the name of the organs included the study in the Organs column refers to the number of traits per organ. For more details on these studies see Appendix S. 114

7.2 Comparison of trait-trait correlations using only gap-filled or only original trait values. Sample sizes (n) of the gap filled and original database are the same to ensure they are comparable. To test whether the gap-filling algorithm had an effect on trait-trait correlation we used a subset of 470769 observations from TRY for five traits (leaf area, SLA, leaf n, plant height and seed mass). First we ran the gap filling algorithm on this dataset. Then using standardize major axis analyses we compared the trait-trait correlations of a dataset only using observed values vs. the exact same observations only using gap filled values. The sample sizes for these trait-trait correlations varied between 1738 for the leaf N-seed mass correlation to 63846 records for the SLA-leaf area correlation. Overall, the difference in the slope value of the trait-trait correlation between the two types of data (gap-filled vs original) ranged from 0.0005 to 0.06, and in the case of the intercepts it varied between 0.006 and 0.11. 115

7.3 Modules of non-woody and woody species across climate regions. The pipe character '|' separates individual modules. Traits across modules may be connected (see Figure 7.3), however they tend to be more connected with other traits within the modules than with traits outside the module. 115

7.4 Trait connections that are robust (i.e. common across groups) across growth forms and climate regions and proposed mechanisms that maintain this connection. The range of R^2 values observed across growth form, and then by growth form across climate regions in this study is provided from the second to fourth columns. We provided mechanism proposed to maintain these trait connection as well as specific hypothesis about this mechanism (Details column). 116

7.5 Trait-trait correlations (r) and precision matrix values (ω) for all land plants, woody and non-woody species. 117

7.6 Woody species trait-trait correlations (r) and precision matrix values (ω) across climate regions 118

7.7 Non-woody species trait-trait correlations (r) and precision matrix values (ω) across climate regions 119

7.8 Trait centrality (i.e. degree) for woody and non-woody species grouped into five different climate regions. Analyses were ran using (A) mass-based, and (N) area based leaf nutrient content (N, and P). 121
7.9 Connectivity (i.e. Edge density) and modularity of plant trait networks for woody and non-woody species across five different climate regions. Analyses were ran using (A) mass-based, and (N) area based leaf N and P content. 121
7.10 Number of present edges and modularity of plant trait networks and modularity for temperature and precipitation gradients. n refers to the number of species in each category. 123
7.11 Trait centrality (i.e. Degree) for woody angiosperms, non-woody forbs and non-woody monocots across a precipitation and temperature gradient, holding the other climate variable constant (see Section 7.4.6). . . . 123

List of Figures

1.1 TRY db (https://www.try-db.org/): (a) A snapshot of TRY db where rows are plants and columns are traits. Blues denote the missing data. It is almost blue. (b) Spatial coverage of measurement sites for plant traits in the TRY db (blue), and the contributing institutes (red) [101]. 5

1.2 Presenting the graphical structure of θ_1, θ_2 and the change $\delta\theta = \theta_1 - \theta_2$ where blue denotes the common edges between θ_1 and θ_2, red edges are only present in θ_1 and green edges are only present in θ_2. a) In this scenario, both θ_1 and θ_2 are sparse that can be estimated correctly even in the low sample setting. Hence, an indirect approach can efficiently estimate the change $\delta\theta$. b) In this scenario, θ_1 and θ_2 are both dense, but $\delta\theta$ is sparse. Since θ_1 and θ_2 can not be estimated correctly in the low sample setting, a direct estimator is a more efficient and consistent approach to estimate the change. 8

1.3 Illustrating the idea behind generalized direct change estimation. First and second columns are the adjacency matrix for two graphical models at different conditions (θ_1 and θ_2) where blues denotes zero (missing edges). Third columns shows the change between two adjacency matrices ($\delta\theta = \theta_1 - \theta_2$), and last column shows the graphical structure of $\delta\theta$. Each row presents an example of $\delta\theta$ with different structures. In all three scenarios, both θ_1 and θ_2 are pretty dense. First row shows the sparsity structure of $\delta\theta$ (a few edges has been changed). Second row presents the group sparsity structure (the connection of two blocks of nodes has been changed). Last row shows the node perturbation structure (the connections of node 5 to all other nodes has been perturbed). The goal of generalized direct change estimation is to estimate $\delta\theta$ (the third column) under different structure without estimating θ_1 and θ_2. 9

1.4 In this example, the matrix X is a tall matrix. All previous algorithms in the literature [21, 22, 117, 144, 174, 175], require either estimating or sampling both latent matrices U or V. The motivation behind our collapsed Monte Carlo inference is to marginalize the tall matrix U and infer the parameters only based on the smaller matrix V. 11

1.5 BHPMF Schematic and Markov blanket of n^{th} row of $U^{(2)}$, $\mathbf{u}_n^{(2)}$, is shown in the red box. In spite of the size of the model, the Gibbs sampler is efficient since the Markov blanket is small and independent of the number of levels. 12

1.6 Trait-trait interaction as a function of temperature (x-axis) and rainfall (y-axis). Edges represent the conditional dependency between traits (nodes). 13

3.1 First row $\delta\theta^*$ has a sparse structure (L_1 norm) and θ_1^* has 3 disconnected star graphs. Second, third, and forth rows $\delta\theta^*$ has group sparse structure (group sparse norm) where θ_1^* has a random graph structure in second row, scale-free structure in third row, and block structure in forth row. Last row $\delta\theta^*$ has two perturbed norm (Node perturbation) and θ_1^* has a random graph structure. Blacks in heatmaps denotes zeros. ROC curve for different structures show in the last column. Direct approach has a better ROC curve for all structures except with scale-free structure of θ_1^*. 39

4.1 (a,b) ROC curves without projection ($\hat{S}$ need not be positive semi-definite), (c,d) ROC curves with projection ($\hat{S}$ is positive semi-definite) with $n = 200$ and under different missing probabilities ($\delta = 0.1 - 0.3$). By increasing number of observed data (smaller δ), the ROC curve approaches the ROC curve of no-missing data ($\delta = 0$). 52

4.2 ROC curve with $\delta = 0.1, 0.2, 0.3$, $p = 100$, and different number of samples (n). For a fixed value of δ, with increasing number of samples, the higher TP rates is obtained. 53

4.3 ROC curve of mGlasso with $n = 200$ and different missing probabilities. mGlasso has a worse performance on non-Gaussian data compared to DoPinG (Figure 4.1). 54

4.4 Precision and Recall Curve with different δ. DoPinG is significantly better than mGlasso for non-Gaussian data. 55

4.5 The graph discovered by DoPinG and mGlasso. 56

5.1 An illustration of bad proposal distribution in importance sampling. Let $p(x) = h^*(x)g^*(x)/Z_p \propto h(x)g(x)$. Neither $h(x) = h^*(x)/Z_h$ nor $g(x) = g^*(x)/Z_g$ are a good candidate proposal distribution since their modes are far away from the one of $p(x)$. 59

5.2 (a,b) Comparison of different proposal distribution (a) Wishart ($\mathcal{W}$) and (b) Inverse Wishart ($\mathcal{IW}$) for sampling mean of $\mathcal{MGIG}_1(\Psi, \Phi, \nu)$ where Λ^* is the mode of $MGIG$. The blue curves are the proposal distribution defined in [229, 233] which can not recover the mode of the $\mathcal{MGIG}$ distribution. 63

5.3 Illustration of 2-dimensional (a) $\mathcal{MGIG}$ distribution (b-f) and different proposal distributions where (b-e) are the proposal described in this chapter where Λ^* is the mode of $MGIG$ and (f) is the proposal defined in [229, 233]. the proposal distribution defined in [229, 233] (f) can not recover the mode of the $\mathcal{MGIG}$ distribution (a). (g) Density of $\mathcal{MGIG}_2(\Psi, \Phi, \nu)$ for 1000 samples generated by each proposal distribution is calculated. More than 90% of samples generated by the previous proposal distribution in [229, 233] ($\mathcal{IW}(\psi, -2\nu)$) have zero $\mathcal{MGIG}$ density leading to $ESS = 40$. Whereas, the new proposal distribution $IW(23\Lambda^*, 20)$ has the $ESS = 550$ which has a very similar shape to the target $\mathcal{MGIG}$ distribution. 64

5.4 Log loss (LL) of CMC and MCMC for different log loss percentile on different datasets presented in the log scale (δ denotes the missing proportion). CMC consistently achieves lower LL compared to MCMC. LL of MCMC increases exponentially (linearly in log scale) by adding data points with higher log loss. Proposal in [30,31] achieved infinity LL for MovieLens. Empty bar represents infinity LL (e.g. 90% and 100% percentile in (d) 73

5.5 LL of CMC and MCMC for different sample size of MovieLens data in the log scale. LL of both CMC and MCMC is decreasing by adding more samples. LL of MCMC is in magnitude 10 times more than CMC. . . . 74

5.6 Density of CMC and MCMC for several data input on MovieLens data. CMC provide distributions with lower LL compared to MCMC e.g. in (a) LL of MCMC is -Inf whereas LL of CMC is -1.78. 75

6.1 (a) BHPMF and (b) MI-BHPMF schematic at level (ℓ). In spite of the size of the model, the Gibbs sampler is efficient since the Markov blanket is small and independent of the number of levels. MI-BHPMF supports multiple inheritance. 78

6.2 a) RMSE of different BHPMF samplers with increasing number of iterations. Block-wise sampler outperforms others. b) BHPMF for all traits and with the inverse of prediction confidence (Std) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the Std is small (more confident). 86
6.3 a,b,e) Spatial coverage of all observation, the highest and lowest confident group. Trait measurements in China or south Africa are more frequent in the uncertain groups (e). Additional measurements in the densely covered regions like China may improve the accuracy. 88
6.4 Scatter plots for pairs of traits (a) on observed true test data, (b) predicted by HPMF, and (c) predicted by BHPMF. BHPMF and PHMF preserve true trait correlations. 89
6.5 MI-BHPMF for all movies with the inverse of prediction confidence (standard deviation) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the standard deviation is small (more confident). 91
6.6 BHPMF for each of the 13 traits with the inverse of prediction confidence (standard deviation) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the standard deviation is small (more confident). 92
7.1 Hypothetical scenarios when determining the conditional dependency among correlated traits. (a) Observed correlation between trait y and trait x, when the effect of trait z has not been considered (dashed gray lines). (b) Conditional dependence between trait x and trait y even after considering trait z, suggesting dependency between trait x and trait y. (c) Conditional independence between trait y and trait x, once trait z has been considered, suggesting that the correlation between x and y was indirectly mediated through z. 101
7.2 Connections between multiple traits across organs (leaves in greens, stems in browns and seed in red) using mass units for leaf N and P content. (A) all terrestrial plants, (B) non-woody species and (C) woody species. 111

7.3 Connections between traits across organs (leaves in greens, stems in browns and seed in red) for woody (A-E) and non-woody species (F-J) in (Tr) Tropical, (Te) Temperate, (Ar) Arid, (Co) Cold, and (Po) Polar environments. Environment types were derived using the Kppen Climate Zones classification system (Peel et al., 2007; see methods). 112
7.4 Connections between traits across organs (leaves, stems and seed) using area units for leaf N and P content. (A) All terrestraial plants included in this study, (B) No-woody species, (C) Woody species. 120
7.5 Connections between traits across organs (leaves, stems and seed) for woody (A-E) and non-woody species (F-J) across a climate gradient and using area-based measurements of leaf N and P content. The climate regions are Tr Tropical, Te Temperate, Ar Arid, Co Cold, Po Polar. . . 120
7.6 Connections among traits across multiple organs for woody angiosperms across a temperature gradient, holding precipitation between 0-500 mm. 122
7.7 Connections among traits across multiple organs for forbs across a temperature gradient, holding precipitation constant between 0-500 mm. . . 122
7.8 Connections among traits across multiple organs for woody angiosperms across a precipitation gradient, holding temperature between 10-20 C. . 124
7.9 Connections among traits across multiple organs for non-woody (A-C) forbs and (D-E) monocots across a precipitation gradient, holding temperature constant between 10-20 C. 124
7.10 Interaction between variables may depend on latent (or unmeasured) variables. For example, in graph a. Wet Street and Wet Grass are conditionally dependent without observing the Rain variable, but they become conditionally independent after observing the Rain Variable (b). Dash lines present the direct edges that cannot be obtained without considering the unmeasured variables. In another example (c), in the presence of a sprinkler the interaction structure may change to the following graph. 125

Chapter 1

Introduction

Many fields in modern science and engineering such as ecology [64, 71], computational biology [12, 39, 184, 232], astronomy [13, 193], signal processing [30, 47, 204], climate science [59, 178], brain imaging [61, 120], natural language processing [10, 80], and many more involve collecting data sets in which the dimensionality of the data p exceeds the sample size n. For example, in computational biology, usually the expression level of thousands genes for about hundreds of patients are measured. A common problem is then to estimate the gene-gene interaction networks which requires estimating $p =$ thousands $\times$ thousands interactions from only hundreds samples (i.e., $n < p$). In another example, consider plant traits analysis where on average three out of 1000 traits are measured (i.e., more than 95% of trait measurements are missing). A typical goal is to fill the gaps in the plant traits matrix for about 1 million ($1M$) plants which requires estimating $p = 1000 \times 1M$ trait values from only a small fraction of measured traits.

In settings where the number of parameters is large relative to the sample size, the use of well studied classical approaches such as least squares regression is problematic since such methods need $n > p$ to be statistically and/or computationally meaningful. In the high dimensional settings, sparse models, or in general constrained structurally models, are usually preferred due to easier interpretation and more accurate and consistent results. For example, a small number of genes may constitute a signature for disease, very few parameters may be required to specify the correlation structure in a time series, or a sparse collection of geometric constraints might completely specify a molecular configuration. Such low-dimensional structure plays an important role in

making high dimensional problems well-posed.

1.1 Structured Graphical Models

The past decade has seen considerable advances in high-dimensional sparse and structured models which continue to stay effective under 'low sample' settings, even when the number n of samples is smaller than the ambient dimensionality p of the model, i.e., $n \leq p$. Such models includes sparse and structured sparse regression models (e.g., Lasso, group Lasso) [186, 201, 236], matrix completion models under low rank assumptions[38, 104, 150, 166], and sparse structure learning of graphical model [15, 100, 140, 164, 165, 228]. Due to the preliminary success in certain application domains, there have been increasing attempts to apply such sparse/structured models to scientific problems such as in climate science, brain sciences, ecology, and genomics [43, 71, 231, 239, 240].

To control the structure and complexity of the model, often regularization terms have been added to the objective function. For instance, in application to linear models, the Lasso or basis pursuit approach [48, 201] is based on a combination of the least-squares loss with ℓ_1-regularization which has been widely applied for feature selection. Similar approaches have been applied to generalized linear models, resulting in more general (non-quadratic) convex programs with ℓ_1-constraints. Several types of regularization have been used for estimating matrices, including standard ℓ_1-regularization, a wide range of sparse group-structured regularizers, as well as regularization based on the nuclear norm (sum of singular values). In high-dimensional settings, regularization serves two purposes: one statistical and the other computational. Statistically, regularization is essential: it prevents overfitting and allows us to design estimators that exploit latent low-dimensional structure in the data to achieve consistency. From the computational point of view, regularization improves the stability of the problem and often leads to computational gains.

In this thesis, we develop probabilistic models for matrix completion and structure learning of graphical models in high-dimensions and apply those advanced methods in plant trait analysis as an application. In certain settings, we focus on Bayesian graphical models since they can produce uncertainty quantified predictions, such as a distribution

over possible values rather than a point estimate; further, since graphical models are modular, combining different models together based on different sources of information can be conceptually straightforward. In this chapter, we start by a brief overview of structure learning of graphical models and low rank matrix completion models, followed by an introduction to the plant trait analysis. Finally, we present an overview of our contributions and developed models.

1.1.1 Structure Learning of Graphical Models

Probabilistic graphical models provides a mechanism for exploiting structure in complex distributions to describe them compactly, and in a way that allows them to be constructed and utilized effectively. These models use a graph-based representation where the nodes are the random variables in our domain, and the edges correspond to direct probabilistic interactions between them. If there is no edge between two nodes, then the corresponding variables are conditionally independent given all other variables.

Often, we may not know the correct graph structure to use for modeling some collection of random variables. Then, it is natural to seek a good graph structure based on sample data. Learning the structures of graphical models have applications in several domains such as learning the gene-gene (or protein-protein) interactions from the gene expression levels data [78, 121], learning the brain neural connectivity [93, 146], and inferring the interaction network among stock markets [82, 84, 88].

In recent years, considerable effort has been invested in obtaining an accurate estimate of sparse structure of graphical models including learning graph structure of Gaussian graphical models [15, 76, 100, 140, 165, 36, 35, 130, 235], Ising graphical models [164], and multivariate Poisson graphical models [228]. Further, the Gaussian graphical models are generalized to Gaussian copula graphical models which can automatically detect potentially nonlinear but monotonic relationships, and correctly identify the nonparametric correlations [128, 227]. The sparse graph strurcture learning have been extended to handle data with missing values in the data [126, 192, 136, 134, 103, 214], which often occur in real world applications, e.g., drop-outs of sensors in a sensor network or missing measurements of temperature or rain in climate.

Throughout the literature, for the high dimensional regime, the non-asymptotic upper bounds on the estimation error has been extensively studied [15, 100, 140, 164,

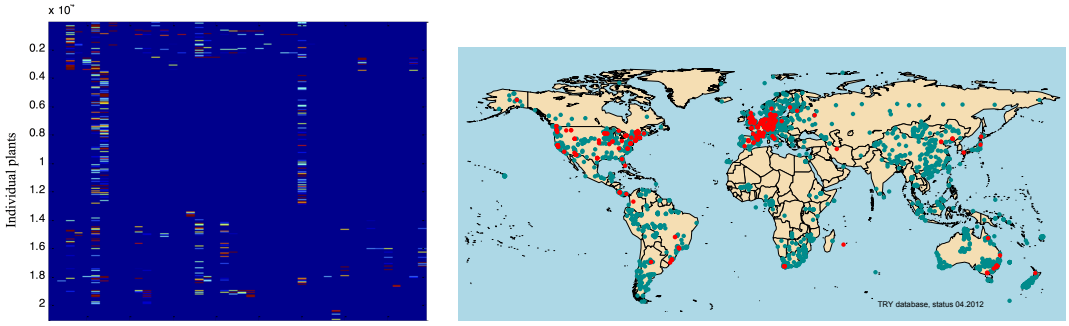
165, 228, 36]. In a high dimensional setting, accurate estimation of the graphical structure depends on how sparse the true graphical structure is. However, the majority of statistical analysis of graphical model estimations assume that all the data points are sampled from the same distribution and provide the sample complexity and convergence rate by considering only one graphical structure for all the observations [36, 164, 165]. New analysis is required to extend the consistency analysis of graphical models when observations draw from multiple distributions.

1.1.2 Low-rank Matrix Completion

Matrix completion is another challenging problem that arise in high dimensional settings. Matrix completion has been extensively studied in the recent literature and have been shown to be successful in a variety of settings [2, 106, 117, 162, 174, 175, 188]. To illustrate the problem statement, consider the collaborative filtering problem of estimating how N users will rate M movies. Clearly, we will have $p = N \times M$ parameters to estimate from n observed ratings. That is, we wish to recover estimates for all possible pairs of movie and user ratings based on only a small fraction of rated films.

The classical statistical setting would require that a small number of entries are missing in order to make an accurate prediction of the ratings. In essence, we would require that majority of users to watch and record a rating for majority of movies, which would be very impractical. However, several modern problems, e.g., recommendation systems, plant traits, work with "mostly missing" matrices where more than 95% data is missing. Intuitively, if we only observe a small fraction of the entries, then there are an infinite number of matrices that can fit the same data observations. In general, there is no way to overcome this unless we impose an implicit structural constraint on the parameter set to reduce the effective size of the parameter space.

In most settings, low rank structural constraint has been imposed to the parameter set, since usually various users (and movies) share similar characteristics, hence users (and movies) can be represented in a low dimensional space. In general, the given sparse matrix $X \in \mathbb{R}^{N \times M}$ is approximated by a low-rank matrix $\hat{X} = UV^T$ where $U \in \mathbb{R}^{N \times D}$ and $V \in \mathbb{R}^{M \times D}$. The latent factors $u_n \in \mathbb{R}^D$, for each row n, and the latent factors $v_m \in \mathbb{R}^D$, for each column m of matrix X are estimated, usually based on alternating optimization [97, 106]. Once the latent factors have been estimated, the inner product



(a) TRY Sparsity (b) TRY spatial coverage

Figure 1.1: TRY db (https://www.try-db.org/): (a) A snapshot of TRY db where rows are plants and columns are traits. Blues denote the missing data. It is almost blue. (b) Spatial coverage of measurement sites for plant traits in the TRY db (blue), and the contributing institutes (red) [101].

of u_n and v_m gives the prediction for the missing entry x_{nm}.

Such methods broadly come in two flavors—from an optimization perspective usually based on a rank or nuclear-norm constraints [176, 225], or, using Bayesian models based on latent factors [174, 175]. Several important variants of such models have been investigated [4, 117, 162, 174, 175, 188], including probabilistic matrix factorization (PMF) and its Bayesian generalizations [174, 175], as well as generalizations to probabilistic tensor factorization [49, 196, 226].

1.2 Plant Trait Analysis

We apply the developed probabilistic structured models to the analysis of plant traits. Plant traits are morphological, anatomical, biochemical, physiological or phenological features of individuals, their component organs or tissues [101]. Examples of traits include the nitrogen content of leaves, leaf area, and plant height. Trait distributions vary across different environmental conditions (e.g., temperature, precipitation, soil moisture), within evolutionary history, and among different species. Understanding trait variation and distribution at local and worldwide spatial scales is an important key to maintain biodiversity and ecosystem functional services (e.g., agricultural and forest productivity, regulation of atmospheric CO_2), and predict the adaptation of planet to human activities and climate change.

To facilitate plant trait analysis, the TRY project (www.try-db.org) was launched in 2007 which brings together different trait databases in a central repository [101]. The TRY database has become the world's largest trait database (covering 1000 traits and 2.1 million plants) and one of the most widely used resources for the ecological community. The TRY database in combination with the recent progress in machine learning provides a unique opportunity to study trait variation and distribution. Nevertheless, we are confronted with two key challenges: (1) Trait sparsity: while unprecedented in coverage, the TRY database is highly sparse (lacking most of the trait information. On average, only three out of 1000 traits are characterized for each individual plant (Figure 1.1(a)). As a result, incorporating the rich information provided by traits in understanding the adaptation of terrestrial ecosystems to climate changes remain difficult; (2) Spatial sparsity: with respect to spatial coverage at global scale, even 2.1 million individual plants provide a sparse coverage (Figure 1.1(b)).

The goal of this thesis is to address the above challenges and analyze plant trait data by proposing novel machine learning models. My main research contributions can be divided into two core components, respectively focusing on developing models (1) to provide trait predictions at individual plant level by incorporating plant taxonomic hierarchy (gap filling) and (2) to provide trait characterization in a given context e.g., climate, soil type, phylogeny, etc., in which they are considered (contextual trait-trait interactions). Understanding trait-trait relationships and trait-environment relationships can help on providing more explicit representation of ecosystem properties, and a more detailed and dynamic representation of trait variation.

TRY database has more than 99% of the entries missing (Figure 1.1(a)). At a high level, the data are similar to that in a recommendation system, with plants corresponding to users and plant traits corresponding to items, e.g., movies. Thus, for the plant-trait gap filling problem, one can use a suitable low-rank model, such probabilistic matrix factorization (PMF) and variants [5, 174, 175, 241]. The performance of models such as PMF on such plant trait gap filling problems [183] is sobering! To understand the reason, note that plants belong to species, and for gap-filling one can simply use the species mean for that trait, e.g., species mean for SLA, leaf N, leaf C, etc. The species mean sets an extremely competitive baseline for any model to beat, and in fact

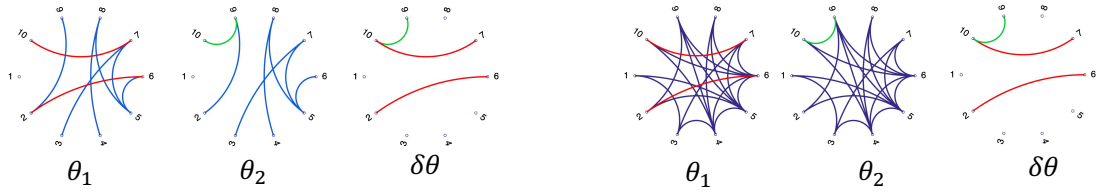
substantially outperforms models such as PMF [71, 179, 183]. From a scientific perspective, the species mean does not constitute an interesting prediction, since it entirely misses out on within species variation of plant traits, the so-called intra-specific variability [6, 53]. Understanding such intra-specific variability, i.e., how plants adjust their traits under different environmental/climatic conditions, is of great scientific interest, and holds critical clues regarding the adaptability of the terrestrial ecosystem under a changing climate.

1.3 Overview and Contributions

In the first half of this thesis, we study structure learning of graphical models. Here, we are interested in learning the graphical structure at different contexts and identify how the graphical structure is evolving under different conditions. For instance, identifying how gene-gene interactions changed from healthy to cancer tissues, learning the changes between brain connectivity in normal and Alzheimer's patients, or learning the changes in the stock market dependency structures. Next, we develop a novel model to estimate the structure of graphical models in presence of missing data and provide the statistical recovery analysis for the proposed estimator.

In the second part of this thesis, we focus on low rank matrix completion problems. In particular, we provide a novel Monte Carlo inference for low rank matrix factorization such as probabilistic matrix factorization (PMF) and Bayesian principle component analysis (BPCA), and proposed a novel Bayesian hierarchical PMF (BHPMF) model that incorporate hierarchical side information.

Finally, we consider applications of probabilistic structured models to plant trait analysis. We apply BHPMF model to fill the gaps in TRY database. The BHPMF model is the-state-of-the-art model for plant trait prediction and is getting increasing visibility and usage in the plant trait analysis [64]. Next, we study the trait-trait interaction structure at different climate zones for different plant growth forms. We briefly explain our contribution for each task in below.



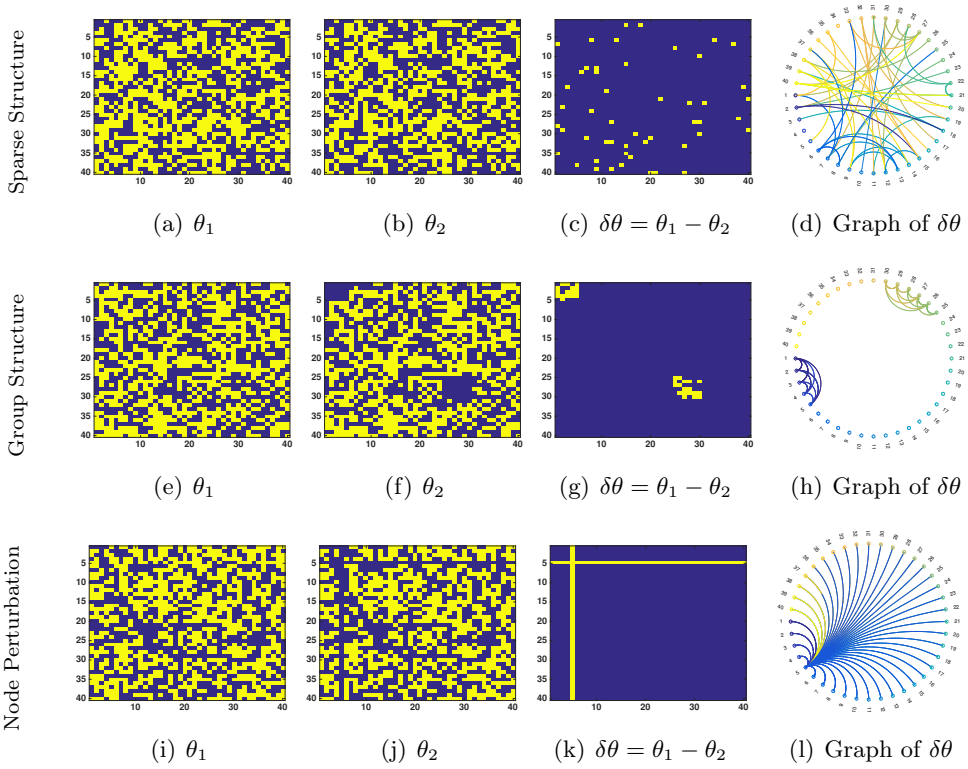
(a) (b)

Figure 1.2: Presenting the graphical structure of θ_1, θ_2 and the change $\delta\theta = \theta_1 - \theta_2$ where blue denotes the common edges between θ_1 and θ_2, red edges are only present in θ_1 and green edges are only present in θ_2. a) In this scenario, both θ_1 and θ_2 are sparse that can be estimated correctly even in the low sample setting. Hence, an indirect approach can efficiently estimate the change $\delta\theta$. b) In this scenario, θ_1 and θ_2 are both dense, but $\delta\theta$ is sparse. Since θ_1 and θ_2 can not be estimated correctly in the low sample setting, a direct estimator is a more efficient and consistent approach to estimate the change.

1.3.1 Generalized Direct Change Estimation

While structure learning in graphical models has been widely studied over the past decade, we focus on the problem of *estimating changes in graphical model structure*: given two sets of samples $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$ respectively drawn from two p-dimensional graphical models with true parameters θ_1^* and θ_2^*, where $\theta_1^*, \theta_2^* \in \mathbb{R}^{p\times p}$, the goal is to estimate the change $\delta\theta^* = (\theta_1^* - \theta_2^*)$. In particular, we focus on the situation when the change $\delta\theta^*$ has structure, such as sparsity, block sparsity, or node-perturbed sparsity, which can be characterized by a suitable (atomic) norm [42, 145]. However, the individual model parameters θ_1^*, θ_2^* need not have any specific structure, and they may both correspond to dense matrices. The goal is to get an estimate $\delta\hat{\theta}$ of the change $\delta\theta^*$ such that the estimation error $\Delta = (\delta\hat{\theta} - \delta\theta^*)$ is small. Such change estimation has potentially wide range of applications including identifying the changes in the neural connectivity networks, the difference between plant trait interactions at different climate conditions, and the changes in the stock market dependency structures.

One can consider two broad approaches for solving such change estimation problems: (i) *indirect change estimation*, where we estimate $\widehat{\theta}_1$ and $\widehat{\theta}_2$ from two sets of samples separately and obtain $\delta\widehat{\theta} = (\widehat{\theta}_1 - \widehat{\theta}_2)$, or (ii) *direct change estimation*, where we directly estimate $\delta\hat{\theta}$ using the two sets of samples, without estimating θ_1 and θ_2 individually.



(a) θ_1 (b) θ_2 (c) $\delta\theta = \theta_1 - \theta_2$ (d) Graph of $\delta\theta$

(e) θ_1 (f) θ_2 (g) $\delta\theta = \theta_1 - \theta_2$ (h) Graph of $\delta\theta$

(i) θ_1 (j) θ_2 (k) $\delta\theta = \theta_1 - \theta_2$ (l) Graph of $\delta\theta$

Figure 1.3: Illustrating the idea behind generalized direct change estimation. First and second columns are the adjacency matrix for two graphical models at different conditions (θ_1 and θ_2) where blues denotes zero (missing edges). Third columns shows the change between two adjacency matrices ($\delta\theta = \theta_1 - \theta_2$), and last column shows the graphical structure of $\delta\theta$. Each row presents an example of $\delta\theta$ with different structures. In all three scenarios, both θ_1 and θ_2 are pretty dense. First row shows the sparsity structure of $\delta\theta$ (a few edges has been changed). Second row presents the group sparsity structure (the connection of two blocks of nodes has been changed). Last row shows the node perturbation structure (the connections of node 5 to all other nodes has been perturbed). The goal of generalized direct change estimation is to estimate $\delta\theta$ (the third column) under different structure without estimating θ_1 and θ_2.

In a high dimensional setting, recent advances [36, 164, 165] illustrate that accurate estimation of the parameter θ^* of a graphical model depends on how sparse or otherwise structured the true parameter θ^* is. For example, if both θ_1^* and θ_2^* are sparse and the samples n_1, n_2 are sufficient to estimate them accurately [164], indirect estimation of $\delta\hat{\theta}$ should be accurate (Figure 1.2(a)). However, if the individual parameters θ_1^* and θ_2^* are somewhat dense, and the change $\delta\theta^*$ has considerably more structure, such as block

sparsity (only a small block has changed) or node perturbation sparsity (only edges from a few nodes have changed) [145], direct estimation may be considerably more efficient both in terms of the number of samples required as well as the computation time (Figures 1.2(b) and 1.3).

Our Contributions: We consider general structured direct change estimation, while allowing the change to have any structure which can be captured by a suitable (atomic) norm $R(\cdot)$ [69]. Our work is a considerable generalization of the existing literature which can only handle sparse changes, captured by the ℓ_1 norm [132]. In particular, our work now enables estimators for more general structures such as group/block sparsity, hierarchical group/block sparsity, node perturbation based sparsity, and so on [14, 42, 145, 151]. The regularized estimator we analyze is broadly a Lasso-type estimator, with key important differences: the objective *does not* decompose additively over the samples, and the objective depends on samples from two distributions.

1.3.2 Gaussian Copula Precision Estimation with Missing Values

Recently, sparse gaussian graphical model structure estimators have also been generalized to handle data with missing values [126, 192, 136, 134, 103], which often occur in real world applications, e.g., drop-outs of sensors in a sensor network or missing measurements of temperature or rain in climate. However, these sparse precision estimators rely on the Gaussian assumption, which may not be appropriate for non-Gaussian datasets. To deal with non-Gaussian data, H. Liu et al. [128] and L. Xue et al. [227] proposed Gaussian copula graphical models where existing estimators can be generalized to the *non-paranormal* distributions simply using one additional procedure, i.e., estimating nonparametric correlations. It has been shown that the nonparanormal is equivalent to Gaussian copula distribution [129, 206, 205]. Therefore, the estimated correlation matrix of the data after transformation can be plugged into the standard sparse graphical structure estimators with Gaussian assumption. The plug-in procedure can leverage existing theoretical results and achieve the optimal statistical rate of convergence for fully observed data.

Our Contributions: In a joint work with Huahua Wang, Soumyadeep Chatterjee and Arindam Banerjee, we generalize the Gaussian copula estimators to handle data with missing values [214]. In particular, our estimator uses two plugin procedures

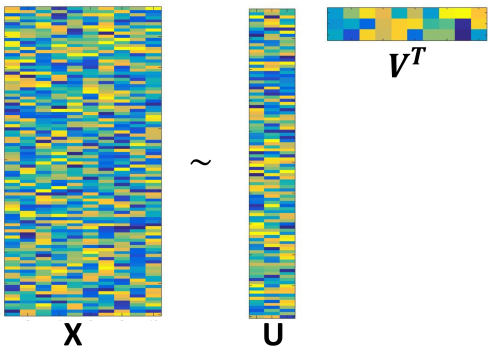


Figure 1.4: In this example, the matrix X is a tall matrix. All previous algorithms in the literature [21, 22, 117, 144, 174, 175], require either estimating or sampling both latent matrices U or V. The motivation behind our collapsed Monte Carlo inference is to marginalize the tall matrix U and infer the parameters only based on the smaller matrix V.

and consists of three steps: (1) estimate nonparametric correlations based on observed values, including Kendalls tau and Spearmans rho; (2) estimate the non-paranormal correlation matrix; (3) plug into existing sparse precision estimators. We show that the consistency rate of our copula estimators depends on the probability of missing values. Through experimental results, we illustrate the effect of sample size and percentage of missing data on the model performance. Experimental results show that our estimator is significantly better than Gaussian estimators.

1.3.3 Collapsed Monte Carlo Inference for Matrix Completion

In the second part of the thesis, we focus on the low rank matrix factorization models such as Probabilistic Matrix Factorization (PMF) or Bayesian Probabilistic Component Analysis (BPCA). For such models, the literature has considered approximate inference methods, such as variational inference [22], gradient descent optimization [117], MCMC [175], Laplace approximation [21, 144], or alternating optimization over U and V [174]. However, all the above inference methods require either estimating or sampling both latent matrices U or V which might not be efficient in several applications especially for tall or fat matrices (i.e., $M \ll N$ or $N \ll M$, figure 1.4).

Out contributions: We propose an efficient inference algorithm for such low rank models [70]. In particular, we show that after analytically marginalizing one of the latent matrices in PMF (or BPCA), the posterior over the other matrix has the Matrix Generalized Inverse Gaussian ($\mathcal{MGIG}$) distribution. We illustrate that the $\mathcal{MGIG}$ distribution is unimodal where the mode can be obtained by solving an *Algebraic Riccati Equation (ARE) [28].* This illustration yields to a novel Collapsed Monte Carlo (CMC)

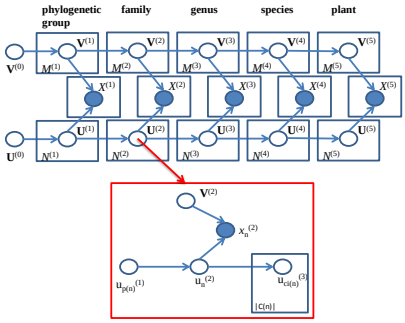


Figure 1.5: BHPMF Schematic and Markov blanket of n^{th} row of $U^{(2)}$, $\mathbf{u}_n^{(2)}$, is shown in the red box. In spite of the size of the model, the Gibbs sampler is efficient since the Markov blanket is small and independent of the number of levels.

inference algorithm for PMF. In particular, we marginalize one of the latent matrices, say U, and propose a direct Monte Carlo sampling from the posterior of the other matrix, say V.

1.3.4 Matrix Completion with Hierarchical Side Information

A key limitation of most matrix factorization models is the inability to use the domain knowledge such as hierarchical side information. In fact, as we illustrate in Section 6.3.4, applying PMF (Probabilistic Matrix Factorization) model [174] which does not incorporate the plant taxonomic hierarchy leads to a performance worse than the simple algorithm MEAN which uses the domain knowledge [183]. Similar hierarchical structure shows up in other applications such as genre (or product type) hierarchy in movies (or products) recommendation. In fact, as we illustrate in Section 6.3.4, applying the PMF model [174] without using the hierarchical information, leads to a performance worse than the simple algorithm MEAN which uses the domain knowledge [183].

Our contribution: The sobering performance of PMF in the plant trait problem led us to take a close look at the domain knowledge available for the problem. An obvious choice was to somehow utilize the plant taxonomic hierarchy, i.e., species, genus, family, etc., in a hierarchical Bayesian low rank model (Figure 1.5). We have developed Bayesian Hierarchical PMF (BHPMF), which use a hierarchy of low rank matrix factorization models, one corresponding to each level of the taxonomic hierarchy, and each level serving as the prior to the next, going all the way down to individual plants [71]. Unlike PMF, the hierarchical low rank models outperformed the species mean baseline substantially, and even captured inter-trait correlations accurately although it was not designed to capture such second order structure explicitly.

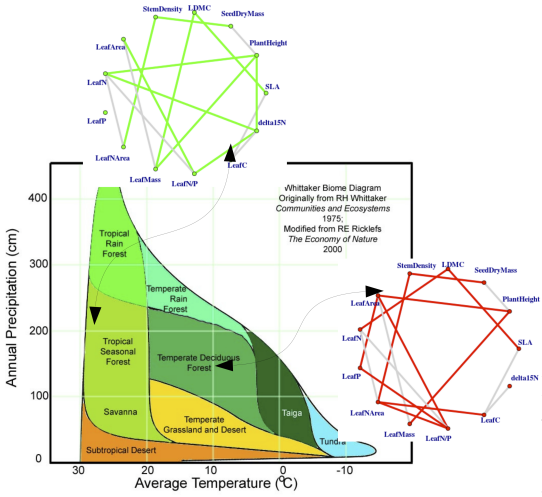


Figure 1.6: Trait-trait interaction as a function of temperature (x-axis) and rainfall (y-axis). Edges represent the conditional dependency between traits (nodes).

1.3.5 Trait-Trait Interactions across Climate Zones

Next, we apply the advances in structure learning of graphical models to estimate trait-trait interactions. Plant traits are not independent of each other, hence an accurate description of their interactions gives us a clearer view of the links between physiological and morphological traits [156, 161], differences between functional groups [155], improves our understanding of the effect of multivariate trait relationships on mechanisms of coexistence [108] and increases accuracy in modeling of ecosystem processes [215]. However, the relationship between traits often depend on the context, e.g., climate, soil type, phylogeny, etc., in which they are considered [3, 170, 169]. For example, two plants/plant types may be close according to phylogeny but in different climates can adapt different trait values.

Our contributions: In a joint work with Habacuc Flores-Moreno, Arindam Banerjee, and others [73], we study plant strategies integration focusing on environmental gradients and growth form (woody and non-woody) to answer our overall question - how does the interaction between traits vary across plant types and environmental gradients? In essence, plants might have different strategies to solve similar environmental dilemmas, and this difference in strategies could also be reflected in the integration among their traits [24]. Here, we first describe the trait-trait interaction network among plants, then we assess how does this trait network change across five broad climate regions (Tropical, Temperate, Arid, Cold and Polar) while accounting for differences in

growth form (woody and non-woody).

Chapter 2

Related Work

2.1 Structured Learning of Graphical Models

Undirected graphical models, also known as Markov Random Fields (MRFs), are important tools for representing multivariate probability distributions which are applied in a wide range of domains such as statistical physics [95], natural language processing [138], image analysis [57] and spatial statistics [172]. The undirected graphical models represent a joint distribution using clique-wise functions over an undirected graph which captures the dependencies among subsets of the p-dimensional discrete random variable $X = (X_1, X_2, \cdots X_p)$. Meaning that feature X_i is conditionally independent of X_j given all other variables if there is no edge in the associated undirected graph structure. The task of graphical model selection is to infer this underlying dependency graph based on data drawn from the corresponding distribution. This task is especially difficult in high-dimensional settings where the number of observations, n is typically even smaller than the number of variables p.

2.1.1 Gaussian Graphical Models

Meinshausen and Buhlmann proposed the neighborhood selection approach and applied the standard Lasso regression on X_j against X_{j^c} to estimate nonzero entries in each row [140]. In the same spirit, Yuan [235] applied the Dantzig selector version of this

regression to estimate Ω column by column. i.e.

$$\min \|\beta\|_1 \qquad \text{s.t.} \frac{1}{n}\|X_{j^c}^T X_j - X_{j^c}^T X_{j^c}\|_\infty \leq \tau. \tag{2.1}$$

Cai et al. [36] further proposed an estimator called CLIME by solving a related optimization problem

$$\hat{\Omega}_n = \underset{\Omega \succ 0}{\operatorname{argmin}} \|\Omega\|_1 \qquad \text{s.t.} \|\hat{\Sigma}_n \Omega - \mathbb{I}\|_\infty \leq \tau. \tag{2.2}$$

In practice, the tuning parameter τ is chosen via cross-validation. However the theoretical choice of $\tau = CM_{n,p} \log p/n$ requires the knowledge of the matrix ℓ_1 norm $M_{n,p} = \|\Omega\|_1$, which is unknown. Later, Cai et al. introduced an adaptive version of CLIME which is data-driven and adaptive to the variability of individual entries of $\hat{\Sigma}_n \Omega - \mathbb{I}$ [35].

Yuan and Lin first proposed to use penalized likelihood methods for estimating sparse precision matrices studied its asymptotic properties for fixed p as $n \to \infty$ [236]. It is easy to see that under the Gaussian assumption the negative log-likelihood up to a constant, can be written as $l\left(X(1), ..., X(n); \Omega\right) = \operatorname{Tr}(\hat{\Sigma}_n \Omega) \log \det(\Omega)$, where $\det(\Omega)$ is the determinant of Ω and $\operatorname{Tr}(.)$ is the trace function. To incorporate the sparsity of Ω, we consider the following penalized log-likelihood estimator with Lasso-type penalty

$$\hat{\Omega}_n = \underset{\Omega \succ 0}{\operatorname{argmin}} \quad \operatorname{Tr}(\hat{\Sigma}_n \Omega) \log \det(\Omega) + \lambda \|\Omega\|_1, \tag{2.3}$$

where $\Omega \succ 0$ means symmetric positive definite.

Rothman et al. analyzed the high-dimensional behavior of estimator (2.3) [173]. Assuming that spectra of Ω are bounded from below and above, the rates of convergence $\sqrt{(p+s)\log p/n}$ and $\sqrt{(1+s)\log p/n}$ under the Frobenius norm and spectral norm are obtained respectively with s being the number of nonzero off-diagonal entries. Lam and Fan studied a generalization of (2.3) and replace the Lasso penalty by general non-convex penalties such as SCAD to overcome the bias issue [112]. Ravikumar et al. applied the primal-dual witness construction to derive the rate of convergence $\sqrt{\log p/n}$ under the sup-norm which in turn leads to convergence rates in the Frobenius and spectral norms as well as support recovery under certain regularity conditions [165].

The results heavily depend on a strong irrepresentability condition imposed on the Hessian matrix $\Gamma = \Sigma \otimes \Sigma$, where $\otimes$ is the tensor (or Kronecker) product. Both sub-Gaussian and polynomial tail cases are considered. However this method cannot be extended to allowing many small nonzero entries .

Although these sparse precision estimators are primarily designed to deal with fully observed data, recently, they have also been generalized to handle data with missing values [126, 192, 136, 134, 103], which often occur in real world applications, e.g., drop-outs of sensors in a sensor network or missing measurements of temperature or rain in climate. To deal with data with missing values, a variety of methods apply expectation maximization (EM) algorithms on imputed data, which are iterative methods but lack theoretical guarantees [126, 192]. Without using the EM algorithm, [134] employed projected gradient descent to solve a sequence of regression problems or PGlasso to estimate the sparse precision matrix of incomplete data. Theoretical guarantees are also established for the PGlasso estimator. M. Kolar and E. Xing introduced a simple plug-in procedure for incomplete data which simply applies existing estimators to the observed data by disregarding the missing values [103]. Such simple plug-in estimators for missing values can leverage existing theoretical results and thus still have similar statistical guarantees, including rate of convergence and consistency. However, these sparse precision estimators rely on the Gaussian assumption, which may not be appropriate for real datasets which are usually non-Gaussian.

To deal with non-Gaussian data, H. Liu et al. [128] proposed Gaussian copula graphical models where existing estimators can be generalized to the *non-paranormal* distributions simply using one additional procedure, i.e., estimating nonparametric correlations. Non-paranormal distributions can be considered as a non-parametric extension of the normal distribution where suitable univariate monotone transformations of the covariates are jointly distributed as a multivariate Gaussian. It has also been shown that the nonparanormal is equivalent to Gaussian copula distribution [129, 206, 205]. Therefore, the estimated correlation matrix of the data after transformation can be plugged into the standard sparse precision estimators with Gaussian assumption. The plug-in procedure can leverage existing theoretical results and achieve the optimal statistical rate of convergence. A similar procedure has also been studied independently by [227].

2.1.2 Ising Graphical Models

In literature, the problem of structure learning for discrete graphical models has attracted considerable attention due to both its importance and difficulty. Score-based approaches through a search procedure generate several candidate graph structures to be scored with a measure of the goodness of fit of the graph. However, the number of graph structures grows super-exponentially, and this problem is in general NP-hard [51]. A complication that arises in graphical model selection with discrete random variables is that the score metrics involve the partition function or cumulant function associated with the Markov random field which is usually computationally intractable [216]. It yields imposing a restricted search space such as directed graphical models [60], trees [52], or hypertrees [190] in the score-based approaches. A method for learning factor graphs based on local conditional entropies and thresholding is proposed in [1] which required the sample complexity of $\Omega(\log p)$, but the computational complexity grows at least as quickly as $O(p^{d+1})$ where d is the maximum degree in the graphical model.

Ravikumar et al. proposed a new model for estimating Ising graphical model structure based on ℓ_1 regularized logistic regression [164]. In particular, the task of recovering of the signed edge vector is reduced to recovering of the signed neighborhood set $N_{\pm}(r)$ for each node r, i.e., capturing both neighborhood structure $N(r)$ and sign pattern for each node r. Given the exponential distribution of the Ising models, the structure of the conditional distribution of X_r given the other variables can be represented as a sigmoid function. Thus, the variable X_r can be viewed as the response variable in a logistic regression in which all of the other variables play the role of the covariates. Later, to impose the sparsity structure, the ℓ_1 regularization is added to the one vs rest logistic regression of X_r on the other variables. The resulting objective function is convex but not differentiable, due to the presence of the ℓ_1-regularizer. By Lagrangian duality, the problem can be re-cast as a constrained problem over the ℓ_1 ball. As a result a minimizer always exists by the Weierstrass theorem. The standard convex programs with an overall computational complexity of order $O(\max\{p, n\}p^3)$ can be applied which is well suited to high-dimensional problems [102]. Their proposed method does not require computing the partition function associated with the Markov random field nor a combinatorial search through the space of graph structures.

Ravikumar et al. provide the theoretical statistical analysis for the formulated estimator and handle the high dimensional setting where both the dimension p and as well as the maximum degree d may tend to infinity as a function of n [164]. They showed that if the Hessian sub-matrix $[\nabla \ell(\theta)]_{SS}$ (i.e., Fisher information matrix) is strictly positive definite, the optimal solution is unique. Under the above conditions, a primal-dual witness procedure is constructed for establishing sufficient conditions for correct signed neighborhood recovery for each node r. Throughout the analysis, it is assumed that the population Fisher information matrix Q^* satisfies the dependency and mutual incoherence conditions. The former by considering the bounded minimum and maximum eigenvalue of Q^*, ensures that the relevant covariates do not become excessively dependent. The later establish that the large number of irrelevant covariates cannot influence the subset of relevant covariates (neighbors of node r). It is shown that under above conditions on the population Hessian matrix, with maximum neighborhood size d and the sample size of $n = \Omega(d^3 \log p)$, for each node r, the ℓ_1-regularized logistic regression, has a unique solution which correctly excludes all edges not in the true neighborhood and can recover all true edgesare not too close to zero (in absolute value). The method can estimate the true graph if minimum edge weight is scaled as $\Omega(\sqrt{\frac{d \log p}{n}})$.

Later, Anandkumar et al. proposed an efficient threshold-based algorithm for structure estimation based on Conditional Mutual Information Thresholding (CMIT) which requires only low order statistics of the data [8]. More specifically, the conditional mutual information test proceeds as follows: one computes the empirical conditional mutual information for each node pair $(i, j) \in V^2$ and finds the conditioning set which achieves the minimum, over all subsets of cardinality at most η. If the above minimum value exceeds the given threshold $\epsilon_{n,p} = \Omega\left(\frac{\log p}{n}\right)$, then the node pair is declared as an edge. Recall that the conditional mutual information is zero iff given X_S, the random variables X_i and X_j are conditionally independent. Thus, the above test seeks to identify non-neighbors, i.e., node pairs which can be separated in the unknown graph G.

The computational complexity of the CMIT algorithm is $O(p^{\eta+2})$. Thus the algorithm is computationally efficient for small η. The parameter η is an upper bound on the size of local vertex-separators in the graph, and is small for many common graph families. Anandkumar et al. show that CMIT is structurally consistent i.e., under bounded

potentials for Ising Models and local-separation property, the CMIT algorithm consistently recovers the structure of the graphical models with probability tending to one. The sample complexity of the CMIT scales as $\Omega\left(J_{min}^{-4}\log p\right)$ and is favorable when the minimum (absolute) edge potential J_{min} is large. This is intuitive since the edges have stronger potentials when J_{min} is large.

2.1.3 Direct Change Estimation

In recent work, Liu et al. [132] proposed a direct change estimator for graphical models based on the ratio of the probability density of the two models [85, 100, 194, 195, 208]. They focused on the special case of L_1 norm, i.e., $\delta\theta^* \in \mathbb{R}^{p^2}$ is sparse, and provided non-asymptotic error bounds for the estimator along with a sample complexity of $n_1 = O(s^2\log p)$ and $n_2 = O(n_1^2)$ for an unbounded density ratio model, where s is the number of the changed edges with p being the number of variables. Liu et al. [133] improved the sample complexity to $\min(n1, n2) = O(s^2\log p)$ when a bounded density ratio model is assumed. Zhao et al. [238] considered estimating direct sparse changes in Gaussian graphical models (GGMs). Their estimator is specific to GGMs and can not be applied to Ising models.

In another work, Zhao et al. [238] estimated the direct changes in Gaussian graphical models by solving a constrained optimization problem and defining the differential graphical model structure as the difference between the two precision matrices (inverse of covariance matrix) at each state. Under the assumption that $\delta\theta^*$ is sparse, Zhao et al. [238] show that the direct estimator is consistent in support recovery and estimation. Their estimator is specific to GGMs and can not be applied to Ising models.

2.2 Low Rank Matrix Completion

Low-rank MF algorithms provide powerful techniques for matrix completion [2, 106, 117, 162, 174, 175, 188]. It has been shown that rank constraint minimization problems can be formulated as trace norm constraints which are convex and can be written as semi-definite constraints [72]. Moreover, Srebro et al. proposed maximum margin matrix factorization as a convex, infinite dimensional alternative to low-rank matrix factorization [191]. Several important variants of low-rank matrix factorization have

been investigated, including PMF [174] and its Bayesian generalization [175, 188] as well as generalizations to probabilistic tensor factorization [2, 196, 226]. A non-linear MF using Gaussian process latent variable models is proposed in [117]. However, one major drawback of the above methods is the inability to incorporate side information.

In order to consider side information, several approaches have been proposed to combine MF with topic modeling [162, 182, 211]. Kernelized PMF was developed to incorporate covariance functions based on kernels over rows and columns in the context of latent factor models for matrix completion [241]. Moreover, probabilistic matrix addition is proposed in [5] to capture covariance structure among rows and among columns at the same time by adding the latent matrices. In a recent work in online advertising [141], hierarchical side information is incorporated into MF in three different ways – hierarchical regularization, agglomerate fitting, and residual fitting. Hierarchical PMF was proposed to incorporate the taxonomic hierarchy into PMF which is the state-of-the-art for plant trait prediction [183].

2.2.1 PMF, PPCA, and Bayesian PCA

Here, we give a review of PMF [174], Probabilistic PCA (PPCA) [202], and Bayesian PCA (BPCA) [21], to illustrate the similarity and differences between the existing ideas and our approach. A related discussion appears in [117]. All these models focus on an (partially) observed data matrix $X \in \mathbb{R}^{N\times M}$. Given latent factors $U \in \mathbb{R}^{N\times D}$ and $V \in \mathbb{R}^{M\times D}$, the rows of X are assumed to be generated according to $\mathbf{x}_{:m} = U\mathbf{v}_m^T + \epsilon$, where $\epsilon \in \mathbb{R}^N$. The different models vary depending on how they handle distributions or estimates of the latent factors U, V. Without loss of generality, for all the analysis through the proposal, we are considering a fat matrix X where $M > N$.

PMF and BPMF: In PMF [174], one assumes independent Gaussian priors for all latent vectors $\mathbf{u}_n$ and $\mathbf{v}_m$, i.e., $\mathbf{u}_n \sim \mathcal{N}(0, \sigma_u^2\mathbb{I}), [n]_1^N$ and $\mathbf{v}_m \sim \mathcal{N}(0, \sigma_v^2\mathbb{I}), [m]_1^M$. Then, one obtains the following posterior over (U, V)

$$p\left(U, V|X, \sigma^2, \sigma_u^2, \sigma_v^2\right) = \prod_{n,m}[\mathcal{N}(x_{nm}\big|\langle\mathbf{u}_n, \mathbf{v}_m\rangle, \sigma^2)]^{\delta_{nm}} \prod_n \mathcal{N}(\mathbf{u}_n\big|0, \sigma_u^2\mathbb{I}) \prod_m \mathcal{N}(\mathbf{v}_{:m}\big|0, \sigma_v^2\mathbb{I}) \ , \tag{2.4}$$

where $\delta_{nm} = 0$ if x_{nm} is missing. PMF obtains point estimates $(\hat{U}, \hat{V})$ by maximizing

the posterior (MAP), based on alternating optimization over U and V [174].

Bayesian PMF (BPMF) [175] considers independent Gaussian priors over latent factors with full covariance matrices, i.e., $\mathbf{u}_n \sim \mathcal{N}(0, \Sigma_u), [n]_1^N$ and $\mathbf{v}_m \sim \mathcal{N}(0, \Sigma_v), [m]_1^M$. Inference is done using Gibbs sampling to approximate the posterior $P(U, V|X)$. At each iteration, U is sampled from the conditional probability of $p(U|V, X)$, followed by sampling V from $p(V|U, X)$ using the updated matrix U at the current iteration.

Probabilistic PCA: In PPCA [202], one assumes independent Gaussian prior over $\mathbf{u}_n$, i.e., $\mathbf{u}_n \sim \mathcal{N}(0, \sigma_u^2 \mathbb{I})$, but V is treated as a parameter to be estimated. In particular, in PPCA, V is chosen so as to maximize the marginalized likelihood of X given by

$$p\left(X \mid V\right) = \int_U p(X|U, V)p(U)dU = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|0, \sigma_u^2 VV^T + \sigma^2 \mathbb{I}). \tag{2.5}$$

Interestingly, as shown in [202], the estimate $\hat{V}$ can be obtained in closed form. For such a fixed $\hat{V}$, the posterior distribution over $U|X, \hat{V}$ can be obtained as:

$$p(U|X, \hat{V}) = \frac{p(X|U, \hat{V})p(U)}{p(X|\hat{V})} = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{u}_n|\Gamma^{-1}\hat{V}^T\mathbf{x}_n, \sigma^{-2}\Gamma\right), \tag{2.6}$$

where $\Gamma = \hat{V}^T\hat{V} + \sigma_u^{-2}\sigma^{-2}\mathbb{I}$. Note that the posterior of the latent factor U in (2.6) depends on both X and $\hat{V}$. For applications of PPCA in visualization, embedding, and data compression, any point x_n in the data space can be summarized by its posterior mean $E[\mathbf{u}_n|\mathbf{x}_n, \hat{V}]$ and covariance $Cov(\mathbf{u}_n|\hat{V})$ in the latent space.

Bayesian PCA: In Bayesian PCA [21], one assumes independent Gaussian priors for all latent vectors $\mathbf{u}_n$ and $\mathbf{v}_m$, i.e., $\mathbf{u}_n \sim \mathcal{N}(0, \sigma_u^2 \mathbb{I})$ and $\mathbf{v}_m \sim \mathcal{N}(0, \sigma_v^2 \mathbb{I})$, $[m]_1^M$. Bayesian posterior inference by Bayes rule considers $p(U, V|X) = p(X|U, V)p(U)p(V)/p(X)$, which includes the intractable partition function

$$p(X) = \int_U \int_V p(X|U, V)p(U)p(V)dUdV \ . \tag{2.7}$$

The literature has considered approximate inference methods, such as variational inference [22], gradient descent optimization [117], MCMC [175], or Laplace approximation [21, 144].

While PPCA and Bayesian PCA were originally considered in the context of embedding and dimensionality reduction, PMF and BPMF have been widely used in the context of matrix completion where the observed matrix X has many missing entries. Nevertheless, as seen from the above exposition, the structure of the models are closely related (also see [117, 116]).

Part I

Structure Learning of Graphical Models

Chapter 3

Generalized Direct Change Estimation in Graphical Models

3.1 Introduction

In this chapter, we consider Ising models and focus on the problem of *estimating changes in Ising model structure*: given two sets of samples $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$ respectively drawn from two p-dimensional Ising models with true parameters θ_1^* and θ_2^*, where $\theta_1^*, \theta_2^* \in \mathbb{R}^{p \times p}$, the goal is to estimate the change $\delta\theta^* = (\theta_1^* - \theta_2^*)$.

We consider general structured direct change estimation, while allowing the change to have any structure which can be captured by a suitable (atomic) norm $R(\cdot)$. Our work is a considerable generalization of the existing literature which can only handle sparse changes, captured by the L_1 norm. In particular, our work now enables estimators for more general structures such as group/block sparsity, hierarchical group/block sparsity, node perturbation based sparsity, and so on [14, 42, 145, 151]. Interestingly, for the unbounded density ratio model, our analysis yields sharper bounds for the special case of ℓ_1 norm, considered by Liu et al. [132]. In particular, when $\delta\theta^*$ is sparse and our estimator is run with L_1 norm, we get a sample complexity of $n_1 = n_2 = O(s \log p)$ which is sharper than $n_1 = O(s^2 \log p)$ and $n_2 = O(n_1^2)$ in [132].

The regularized estimator we analyze is broadly a Lasso-type estimator, with key important differences: the objective *does not* decompose additively over the samples, and the objective depends on samples from two distributions. The estimator builds on the

density ratio estimator in [132], but works with general norm regularization [14, 42, 151] where the regularization parameter λ_{n_1,n_2} depends on the sample size for both Ising models. Our analysis is quite different from the existing literature in change estimation. Liu et al. [132] build on the primal-dual witness approach of Wainwright [210], which is effective for the special case of L_1 norm. Our analysis is largely geometric, where generic chaining [200] plays a key role, and our results are in terms of Gaussian widths of suitable sets associated with the norm [14, 42].

The rest of the chapter is organized as follows. In Section 3.2, we introduce the direct change estimator based on the ratio of the probability density of the Ising models. In Section 3.3, we establish statistical consistency of the direct change estimator.

3.2 Generalized Direct Change Estimation

We consider the following optimization problem

$$\underset{\delta\theta}{\operatorname{argmin}} \;\; \mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) + \lambda_{n_1,n_2} R(\delta\theta), \tag{3.1}$$

where $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$ are two sets of i.i.d binary samples drawn from from Ising graphical models with parameter θ_1^* and θ_2^*, respectively, each $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ are $p-$dimensional vectors, and n_1, n_2 are the respective sample sizes.

In this Section, we first give a brief background on Ising model selection. Then, we explain how to develop the loss function $\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})$ based on the density ratio [85, 100, 194, 208] to directly estimate $\delta\theta = \theta_1 - \theta_2$, and finally we describe how to solve the optimization problem (3.1) for any norm $R(\delta\theta)$.

3.2.1 Ising Model

Let $X = (X_1, X_2, \cdots, X_p)$ denote a random vector in which each variable $X_s \in \{-1, 1\}$. Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \cdots, p\}$ and edge set E whose elements are unordered pairs of distinct vertices. The pairwise Ising Markov random field associated with the graph G over the random vector X is

$$P(X = \mathbf{x}|\theta^*) = \frac{1}{Z(\theta^*)} \exp\{\sum_{s,t\in E} \theta_{s,t}^* x_s x_t\} \tag{3.2}$$

$$= \frac{1}{Z(\theta^*)} \exp\{\langle \theta^*, T(\mathbf{x}) \rangle\} \tag{3.3}$$

$$= \frac{1}{Z(\Theta^*)} \exp\{\mathbf{x}^T \Theta^* \mathbf{x}\} \tag{3.4}$$

where $T(\mathbf{x}) = \{x_s x_t\}_{s,t=1}^{p}$ is a vector of size $m = p^2$, $\theta^* = \{\theta^*_{s,t}\}_{s,t=1}^{p} \in \mathbb{R}^m$ and $\langle ., . \rangle$ is the inner product operator, and $\Theta^* \in \mathbb{R}^{p \times p}$ where $\Theta^*_{s,t} = \theta^*_{s,t}$. Note that basic Ising models also have non-interacting terms like $\alpha_s x_s$ and we are assuming these terms are zero, and they do not affect the dependency structure.

The parameter θ^* associated with the structure of the graph G reveals the statistical conditional independence structure among the variables i.e., if $\theta^*_{s,t} = 0$, then feature X_s is conditionally independent of X_t given all other variables and there is no edge in the graph G.

The partition function, $Z(\theta^*)$, plays the role of a normalizing constant, ensuring that the probabilities add up to one which is defined as

$$Z(\theta^*) = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{\langle \theta^*, T(\mathbf{x}) \rangle\} = \exp\{\Psi(\theta^*)\}, \tag{3.5}$$

where $\mathcal{X}$ be the set of all possible configurations of X.

3.2.2 Loss Function

Here, we build the loss function based on equation (3.3). Similarly, one can rewrite the loss function based on (3.4) if the regularization function is over matrices.

Consider two Ising models with parameters $\theta^*_1 \in \mathbb{R}^{p^2}$ and $\theta^*_2 \in \mathbb{R}^{p^2}$. Following Liu et. al [131, 132], a direct estimate for the changes detection problem based on density ratio can be posed as follows

$$r(X = \mathbf{x}|\delta\theta) = \frac{p(X = \mathbf{x}|\theta_1)}{p(X = \mathbf{x}|\theta_2)} = \underbrace{\frac{\exp\{\langle T(\mathbf{x}), \theta_1 \rangle\}}{\exp\{\langle T(\mathbf{x}), \theta_2 \rangle\}}}_{r^*(\mathbf{x}|\delta\theta)} \underbrace{\frac{Z(\theta_2)}{Z(\theta_1)}}_{1/Z(\delta\theta)}$$

$$= \frac{\exp\{\langle T(\mathbf{x}), \delta\theta \rangle\}}{Z(\delta\theta)}, \tag{3.6}$$

where the parameter $\delta\theta = \theta_1 - \theta_2$ encodes the change between two graphical models θ_1 and θ_2.

First, we show that $Z(\delta\theta) = E_{X\sim p(X|\theta_2)}[e^{\langle T(X),\delta\theta\rangle}]$:

$$\begin{aligned} Z(\delta\theta) &= \frac{Z(\theta_1)}{Z(\theta_2)} = \frac{1}{Z(\theta_2)}\sum_{\mathbf{x}\in\mathcal{X}} e^{\langle T(\mathbf{x}),\theta_1\rangle} \\ &= \frac{1}{Z(\theta_2)}\sum_{\mathbf{x}\in\mathcal{X}} e^{\langle T(\mathbf{x}),\theta_2\rangle}\frac{e^{\langle T(\mathbf{x}),\theta_1\rangle}}{e^{\langle T(\mathbf{x}),\theta_2\rangle}} \\ &= \sum_{\mathbf{x}\in\mathcal{X}} \underbrace{\frac{e^{\langle T(\mathbf{x}),\theta_2\rangle}}{Z(\theta_2)}}_{p(\mathbf{x}|\theta_2)} e^{\langle T(\mathbf{x}),\delta\theta\rangle} = E_{X\sim p(X|\theta_2)}[e^{\langle T(X),\delta\theta\rangle}]. \end{aligned} \tag{3.7}$$

Next, using the samples $\mathfrak{X}_2^{n_2}$ from $p(X|\theta_2)$, we estimate $Z(\delta\theta)$ empirically as

$$\hat{Z}(\delta\theta) = \frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle T(\mathbf{x}_i^2),\delta\theta\rangle\}, \tag{3.8}$$

and the sample approximation of $r(X|\delta\theta)$ is given as

$$\begin{aligned} \hat{r}(X=\mathbf{x}|\delta\theta) &= \frac{r^*(X=\mathbf{x}|\theta_1)}{\hat{Z}(\delta\theta)} \\ &= \frac{\exp\{\langle T(\mathbf{x}),\delta\theta\rangle\}}{\frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle T(\mathbf{x}_i^2),\delta\theta\rangle\}}. \end{aligned} \tag{3.9}$$

Using the fact that $r(X|\delta\theta^*)q(X|\theta_2^*) = p(X|\theta_1^*)$, we approximate $\hat{r}(X|\delta\theta)$, by minimizing the KL divergence,

$$\begin{aligned} &KL\left(p(X|\theta_1^*)\|\hat{r}(X|\delta\theta)p(X|\theta_2^*)\right) \\ &= \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\theta_1^*)\log\frac{p(\mathbf{x}|\theta_1^*)}{p(\mathbf{x}|\theta_2^*)\hat{r}(\mathbf{x}|\delta\theta)} \\ &= \underbrace{KL\left(p(X|\theta_1^*)\|p(X|\theta_2^*)\right)}_{\text{Constant}} - E_{X\sim p(X|\theta_1^*)}\left[\log\hat{r}(X|\delta\theta)\right] \end{aligned} \tag{3.10}$$

Thus, using the samples $\mathfrak{X}_1^{n_1}$ and $\mathfrak{X}_2^{n_2}$, we define the empirical loss function

$$\mathcal{L}(\delta\theta;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2}) = \frac{-1}{n_1}\sum_{i=1}^{n_1}\log\hat{r}(\mathbf{x}_i^1|\delta\theta) \tag{3.11}$$

$$= \frac{-1}{n_1}\sum_{i=1}^{n_1}\langle T(\mathbf{x}_i^1), \delta\theta\rangle + \underbrace{\log \frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle T(\mathbf{x}_i^2), \delta\theta\rangle\}}_{\hat{\Psi}(\delta\theta)}$$

Remark 1 Note that the loss function (3.11) does not additively decompose over the samples. The second term in (3.11) is the logarithm over sum of a function of samples.

3.2.3 Optimization

The optimization problem (3.1) has a composite objective with a smooth convex term corresponding to the loss function (3.11) and a a potentially non-smooth convex term corresponding to the regularizer. In this section, we present an algorithm in the class of Fast Iterative Shrinkage-Thresholding Algorithms (FISTA) for efficiently solving the problem (3.1) [20]. For convenience, we refer the loss function $\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})$ as $\mathcal{L}(\delta\theta)$ and we drop the subscript $\{n_1, n_2\}$ of λ_{n_1,n_2}.

One of the most popular methods for composite objective functions is in the class of FISTA where at each iteration we linearize the smooth term and minimize the quadratic approximation of the form

$$\begin{aligned} Q_L(\delta\theta, \delta\theta_t) :=& \mathcal{L}(\delta\theta) + \langle \delta\theta - \delta\theta_t, \nabla\mathcal{L}(\delta\theta_t)\rangle \\ &+ \frac{L}{2}\|\delta\theta - \delta\theta_t\|_2^2 + \lambda R(\delta\theta), \end{aligned} \tag{3.12}$$

where L denotes the Lipschitz constant of the loss function $\mathcal{L}(\delta\theta)$. Ignoring constant terms in $\delta\theta_t$, the unique minimizer of the above expression (3.12) can be written as

$$\begin{aligned} p_L(\delta\theta_t) &= \arg\min_{\delta\theta} Q_{\mathcal{L}}(\delta\theta, \delta\theta_t) \\ &= \arg\min_{\delta\theta} \lambda R(\delta\theta) + \frac{L}{2}\left\|\delta\theta - \left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)\right\|_2^2 \\ &= \arg\min_{\delta\theta} \frac{\lambda}{L} R(\delta\theta) + \frac{1}{2}\left\|\delta\theta - \left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)\right\|_2^2. \end{aligned} \tag{3.13}$$

In fact, the updates of $\delta\theta$ is to compute certain proximal operators of the non-smooth term $R(.)$. In general, the proximal operator $\text{prox}_h(\mathbf{x})$ of a closed proper convex function

$h : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is defined as

$$\text{prox}_h(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left(h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right). \tag{3.14}$$

Thus, the unique minimizer (3.13) correspond to $\text{prox}_{\frac{\lambda}{L}R}\left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)$ which has rate of convergence of $O(1/t)$ [152, 157].

To improve the rate of convergence, we adapt the idea of FISTA algorithm. The main idea is to iteratively consider the proximal operator prox(.) at a specific linear combination of the previous two iterates $\{\delta\theta_t, \delta\theta_{t-1}\}$

$$\xi_{t+1} = \delta\theta_t + \alpha_{t+1}\left(\delta\theta_t - \delta\theta_{t-1}\right), \tag{3.15}$$

instead of just the previous iterate $\delta\theta_t$. The choice of α_{t+1} follows Nesterovs accelerated gradient descent [152, 157] and is detailed in Algorithm 1. The iterative algorithm simply updates

$$\delta\theta_{t+1} = \text{prox}_{\frac{\lambda}{L}R}\left(\xi_{t+1} - \frac{1}{L}\nabla\mathcal{L}(\xi_{t+1})\right). \tag{3.16}$$

The algorithm has a rate of convergence of $O(1/t^2)$ [20].

3.2.4 Regularization Function

We assume that the optimal $\delta\theta^*$ is sparse or suitably 'structured' where such structure can be characterized by having a low value according to a suitable norm $R(\delta\theta^*)$. In below, we provide a few examples of such a norm.

L_1 **norm:** One example for $R(.)$ we will consider throughout the chapter is the L_1 norm regularization. We use L_1 norm if only a few edges has changed (1st row in Figure 1.3). In particular, we consider $R(\delta\theta) = \|\delta\theta\|_1$ if number of non-zeros entries in $\delta\theta^*$ is $s < p^2$. The $\text{prox}_{\frac{\lambda}{L}\|.\|_1}(.)$ is given by the elementwise soft-thresholding operation [187] as

$$\left[\text{prox}_{\frac{\lambda}{L}\|.\|_1}\right]_i(\mathbf{z}) = \text{sign}(\mathbf{z}_i).\max(0, \mathbf{z}_i - \frac{\lambda}{L}). \tag{3.21}$$

Algorithm 1 Generalized Direct Change Estimator

1: **Input:** $L_0 > 0$, $\mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}$
2: **Step 0.** Set $\xi_1 = \delta\theta_0$, $t = 1$
3: **Step** t**.** ($t \geq 1$) Find the smallest non-negative integers i_t such that with $\tilde{L} = 2^{i_t} L_{t-1}$

$$\mathcal{L}\left(p_{\tilde{L}}(\xi_t)\right) + R\left(p_{\tilde{L}}(\xi_t)\right) \leq Q_{\tilde{L}}\left(p_{\tilde{L}}(\xi_t), \xi_t\right). \tag{3.17}$$

4: Set $L_t = 2^{i_t} L_{t-1}$ and Compute

$$\delta\theta_t = \text{prox}_{\frac{\lambda}{L}R}\left(\xi_t - \frac{1}{L}\nabla\mathcal{L}(\xi_t)\right) \tag{3.18}$$

$$\beta_{t+1} = \frac{1 + \sqrt{1 + 4\beta_t^2}}{2} \tag{3.19}$$

$$\xi_{t+1} = \delta\theta_t + \left(\frac{\beta_t - 1}{\beta_{t+1}}\right)(\delta\theta_t - \delta\theta_{t-1}) \tag{3.20}$$

Group-sparse norm: Another popular example we consider is the group-sparse norm. We use group lasso norm if a group of edges has changed (2nd row in Figure 1.3). For some kinds of data, it is reasonable to assume that the variables can be clustered (or grouped) into types, which share similar connectivity or correlation patterns. Let $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{N_G}\}$ denote a collection of groups, which are subsets of variables. We assume that $\delta\Theta^*(s,t) = 0$ for any variable $s \in G_g$ and for any variable $t \in G_h$. In the group sparse setting for any subset $S_\mathcal{G} \subseteq \{1, 2, \cdots N_G\}$ with cardinality $|S_\mathcal{G}| = s_\mathcal{G}$, we assume that the parameter $\delta\Theta^*$ satisfies $\{\delta\Theta^*_{s,t} = 0 : s,t \in G_g \;\&\; g \notin S_\mathcal{G}\}$. We will focus on the case when $R(\delta\Theta) = \sum_{g=1}^{N_G} \|\delta\Theta(s,t) : s,t \in G_g\|_F$ [139]. Let $\delta\Theta_{G_g}$ bd the sub-matrix of $\delta\Theta$ covering nodes in G_g. Proximal operator is given by the group specific soft-thresholding operation.

$$\left[\text{prox}_{\frac{\lambda}{L}R}\right]_g(\delta\Theta) = \frac{\max(\|\delta\Theta_{G_g}\|_F - \frac{\lambda}{L}, 0)}{\|\delta\Theta_{G_g}\|_F}. \tag{3.22}$$

Node perturbation: Another example is the row-column overlap norm (RCON) [145] to capture perturbed nodes i.e., nodes that have a completely different connectivity pattern to other nodes among two networks (3rd row in Figure 1.3). A special case of RCON we are interested is $\sum_{i=1}^{p} \|V_i\|_q$ where $\delta\Theta = V + V^T$, and V_i is the $i-$th column of matrix V. This norm can be viewed as overlapping group lasso [145] and thus can be solved by

applying Algorithm 1 with proximal operator for overlapping group lasso [234]. Also, we can write problem (3.1) as a constrained optimization

$$\begin{aligned} \underset{\delta\Theta, V}{\text{argmin}} \quad & \mathcal{L}(\delta\Theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) + \lambda_1 \delta\Theta\|_1 + \lambda_{n_1,n_2} \sum_{i=1}^{p} \|V_i\|_q \\ \text{s.t} \quad & \delta\Theta = V + V^T, \end{aligned} \tag{3.23}$$

and solve it by applying in-exact ADMM techniques [145].

3.3 Theoretical Analysis

Our goal is to provide non-asymptotic bounds on $\|\Delta\|_2 = \|\delta\theta^* - \delta\hat{\theta}\|_2$ between the true parameter $\delta\theta^*$ and the minimizer $\delta\hat{\theta}$ of (3.1). In this section, we describe various aspects of the problem, introducing notations along the way, and highlight our main result.

3.3.1 Background and Assumption

Gaussian Width: In several of our proofs, we use the concept of Gaussian width [42, 83], which is defined as follows.

Definition 3.3.1 *For any set $A \in \mathbb{R}^p$, the* Gaussian width *of the set A is defined as:*

$$w(A) = E_g \left[\sup_{u \in A} \langle g, u \rangle \right] . \tag{3.24}$$

where the expectation is over $g \sim N(0, \mathbb{I}_{p\times p})$, a vector of independent zero-mean unit-variance Gaussian random variable.

The Gaussian width $w(A)$ provides a geometric characterization of the size of the set A. Consider the Gaussian process $\{Z_u\}$ where the constituent Gaussian random variables $Z_u = \langle u, g \rangle$ are indexed by $u \in A$, and $g \sim N(0, \mathbb{I}_{p\times p})$. Then the Gaussian width $w(A)$ can be viewed as the expectation of the supremum of the Gaussian process $\{Z_u\}$. Bounds on the expectations of Gaussian and other empirical processes have been widely studied in the literature, and we will make use of generic chaining for some of our analysis [25, 118, 199, 200].

The Error Set: Consider solving the problem (3.1), under assumption $\lambda_{n_1,n_2} > \beta R^* \left(\nabla \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)$, where $\beta > 1$ and $R^*(.)$ is the dual norm of $R(.)$. Banerjee et al. [14] show that for any convex loss function the error vector $\Delta = (\delta\theta^* - \delta\hat{\theta})$ lies in a restricted set that is characterized as

$$\begin{aligned} E_r &= E_r(\delta\theta^*, \beta) \\ &= \left\{ \Delta \in \mathbb{R}^p \;\middle|\; R(\delta\theta^* + \Delta) \leq R(\delta\theta^*) + \frac{1}{\beta} R(\Delta) \right\} . \end{aligned} \tag{3.25}$$

Restricted Strong Convexity (RSC) Condition: The sample complexity of the problem (3.1) depends on the RSC condition [151], which ensures that the estimation problem is strongly convex in the neighborhood of the optimal parameter [14, 151]. A convex loss function satisfies the RSC condition in $C_r = cone(E_r)$, i.e., $\forall \Delta \in C_r$, if there exists a suitable constant κ such that

$$\begin{aligned} \delta\mathcal{L}(\delta\theta^*, u) :=& \mathcal{L}(\delta\theta^* + u) - \mathcal{L}(\delta\theta^*) - \langle \nabla\mathcal{L}(\delta\theta^*), u \rangle \\ \geq& \kappa \|u\|_2^2 \end{aligned} \tag{3.26}$$

Deterministic Recovery Bounds: If the RSC condition is satisfied on the error set C_r and λ_{n_1,n_2} satisfies the assumptions stated earlier, for any norm $R(.)$, Banerjee et al. [14] show a deterministic upper bound for $\|\Delta\|_2$ in terms of λ_{n_1,n_2}, κ, and the norm compatibility constant $\Psi(C_r) = \sup_{\mathbf{u} \in C_r} \frac{R(\mathbf{u})}{\|\mathbf{u}\|_2}$, as

$$\|\Delta\|_2 \leq \frac{1+\beta}{\beta} \frac{\lambda_{n_1,n_2}}{\kappa} \Psi(C_r) . \tag{3.27}$$

Smooth Density Ratio Model Assumption: For any vector $\mathbf{u}$ such that $\|\mathbf{u}\|_2 \leq \|\delta\theta^*\|_2$ and every $\epsilon \in R$, the following inequality holds:

$$E_{X \sim p(X|\theta_2)}[\exp\{\epsilon \; r(X|\delta\theta^* + \mathbf{u}) - 1\}] \leq \exp\{\epsilon^2\}.$$

A similar assumption is used in the analysis of Liu et al. [132].

Remark 2 Bounded density ratio is a special case satisfying the smooth density ratio assumption. Lemma 1 shows a sufficient condition under which the density ratio is

bounded.

Lemma 1 *Consider two Ising Model with true parameters θ_1^* and θ_2^*. Let $d_1, d_2 \gg s$ where $\|\theta_1^*\|_0 = d_1$, $\|\theta_2^*\|_0 = d_2$, and $\|\delta\theta^*\|_0 = s$. Assume*

$$\min_{i,j=1\cdots p}(|\theta_1^*(i,j)|) \geq \frac{1}{d_1 - 1} - \frac{c_1}{(d_1-1)s} \tag{3.28}$$

$$\min_{i,j=1\cdots p}(|\theta_2^*(i,j)|) \geq \frac{1}{d_2 - 1} - \frac{c_2}{(d_2-1)s}, \tag{3.29}$$

where c_1 and c_2 are positive constants. Then the density ratio $r(X = \mathbf{x}|\delta\theta^*)$ is bounded.

Note that if individual graphs are dense, then the conditions (3.28) and (3.29) are satisfied and as a result the smooth density ratio is satisfied.

Remark 3 In this chapter, we focus on the Ising graphical model. But, our statistical analysis holds for any graphical models that satisfy the above mentioned assumption. Through our analysis, no assumption is required on the individual graphical models.

3.3.2 Bounds on the regularization parameter

To get the recovery bound (3.27) above, one needs to have $\lambda_{n_1,n_2} \geq \beta R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)$. However, the bound on λ_{n_1,n_2} depends on unknown quantity $\delta\theta^*$ and the samples $\mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}$ and is hence random. To overcome the above challenges, one can bound the expectation $E[R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)]$ over all samples of size n_1 and n_2, and obtain high-probability deviation bounds. The goal is to provide a sharp bound on λ_{n_1,n_2} since the error bound in (3.27) is directly proportional to λ_{n_1,n_2}.

In theorem 2, we characterize the expectation $E[R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)]$ in terms of the Gaussian width of the unit norm-ball of $R(.)$, which leads to a sharp bound. The upper bound on Gaussian width of the unit norm-ball of R for atomic norms which covers a wide range of norms is provided in [42, 46].

Theorem 2 *Define $\Omega_R = \{u : R(u) \leq 1\}$. Let $\phi(R) = \sup_{\mathbf{u}} \frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}$. Assume that for any $\mathbf{u}$ that $\|\mathbf{u}\| \leq \|\theta^*\|$*

$$\frac{1}{2}\lambda_{\max}\left(\nabla^2\mathcal{L}(\delta\theta^* + \mathbf{u})\right) \leq \eta_0, \tag{3.30}$$

where $\lambda_{\max}(.)$ *is the maximum eigenvalue. Then under the smooth density ratio assumption, we have*

$$E\left[R^*(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2}))\right] \leq \frac{2\sqrt{\eta_0}(c_1 w\left(\Omega_R\right)+\phi(R))}{\sqrt{\min(n_1,n_2)}}.$$

and with probability at least $1-c_2e^{-\epsilon^2}$

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right) \leq \frac{c_2(1+\epsilon)w(\Omega_R)+\tau_1}{\sqrt{\min(n_1,n_2)}}.$$

where c_1 *and* c_2 *are positive constants,* $\tau_1 = 2\sqrt{\eta_0}\phi(R)$*, and* $w(\Omega_R)$ *is the Gaussian width of set* Ω_R*.*

Note, that our analysis hold for any norm and it is expressed in terms of the Gaussian width. In the following, we give the bound on the regularization parameter for two examples of the regularization function $R(.)$.

Corollary 3 *If* $R(\delta\theta)$ *is the* L_1 *norm, and* $\delta\theta \in \mathbb{R}^{p^2}$ *then with high probability we have the bound*

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right) \leq \frac{\eta_2\sqrt{\log p}}{\sqrt{\min(n_1,n_2)}}. \tag{3.31}$$

Corollary 4 *If* $R(\delta\theta)$ *is the group-sparse norm, and* $\delta\theta \in \mathbb{R}^{p^2}$ *then with high probability we have the bound*

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right) \leq \frac{\eta_2\sqrt{m+\log N_G}}{\sqrt{\min(n_1,n_2)}}, \tag{3.32}$$

where $\mathcal{G} = \{\mathcal{G}_1,\cdots,\mathcal{G}_{N_G}\}$ *is a collection of groups,* $m = \max_i |\mathcal{G}_i|$ *is the maximum size of any group.*

3.3.3 RSC Condition

In this Section, we establish the RSC condition for direct change detection estimator (3.1). Simplifying the expression and applying mean value theorem twice on the left side of RSC condition (3.26), for $\forall\gamma_i \in [0,1]$, we have

$$\delta\mathcal{L}(\delta\theta^*,u) := \mathcal{L}(\delta\theta^*+u) - \mathcal{L}(\delta\theta^*) - \langle\nabla\mathcal{L}(\delta\theta^*),u\rangle$$

$$\geq u^T \nabla^2 \mathcal{L}(\delta\theta^* + \gamma_i u)u. \tag{3.33}$$

Thus, the RSC condition depends on the non-linear terms of loss function. Recall that the nonlinear term, second term, in Loss function (3.1) which is the approximation of the log-partition functions only depends on n_2. As a results, only samples of $\mathfrak{X}_2^{n_2}$ affect the RSC conditions. Our analysis is an extension of the results on [14] using the generic chaining. We show that, with high probability the RSC condition is satisfied once samples n_2 crosses $w^2(C_r \cap S^{d-1})$ the Gaussian width of restricted error set. The bound on Gaussian width of the error set for atomic norms has been provided in [46].

Let $r_i = r(X = \mathbf{x_i^2}|\delta\theta^*)$ and $\bar{\varepsilon}$ denote the probability that r_i exceeds some constant T: $\bar{\varepsilon} = p(r_i > T) \leq 2e^{-\frac{T^2}{2}}$.

Theorem 5 *Let $X \in \mathbb{R}^{n\times p}$ be a design matrix with independent isotropic sub-Gaussian rows with $\|\|X_i\|\|_{\Psi_2} \leq \kappa$. Then, for any set $A \subseteq S^{p-1}$, for suitable constants η, c_1, $c_2 > 0$ with probability at least $1 - \exp(-\eta w^2(A))$, we have*

$$\inf_{u\in A} \partial\mathcal{L}(\theta^*; u, X) \geq c_1 \underline{\rho}^2 \left(1 - c_2\kappa_1^2 \frac{w(A)}{\sqrt{n_2}}\right) - \tau \tag{3.34}$$

where $\kappa_1 = \frac{\kappa}{\bar{\varepsilon}}$, $\underline{\rho}^2 = \inf_{\mathbf{u}\in A} \rho_{\mathbf{u}}^2$ with $\rho_{\mathbf{u}}^2 = E\left[\left\langle \mathbf{u}, T(X_i^2)\right\rangle^2 \mathbb{I}(r_i > T)\right]$, and τ is smaller than the first term in right hand side. Thus, for $n_2 \geq c_2 w^2(A)$, with probability at least $1 - \exp(-\eta w^2(A))$, we have $\inf_{u\in A} \partial\mathcal{L}(\theta^*; u, X) > 0$.

3.3.4 Statistical Recovery

With the above results in place, from (3.27), Theorem 6 provides the main recovery bound for generalized direct change estimator (3.1).

Theorem 6 *Consider two set of i.i.d samples $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$. Define $\Omega_R = \{u : R(u) \leq 1\}$. Assume that $\delta\hat{\theta}$ is the minimizer of the problem (3.1). Then, with probability at least $1 - \eta_0 e^{-\epsilon^2}$ the followings hold*

$$\lambda_{n_1,n_2} \geq \frac{\eta_1}{\sqrt{\min(n_1, n_2)}}(w(\Omega_R) + \epsilon) \tag{3.35}$$

and for $n_2 \geq cw^2(C_r \cap S^{d-1})$*, with high probability, the estimate* $\delta\hat{\theta}$ *satisfies*

$$\|\Delta\|_2 \leq O\left(\frac{w(\Omega_R)}{\sqrt{\min(n_1, n_2)}}\right)\Psi(C_r)\ , \tag{3.36}$$

where $w(.)$ *is the Gaussian width of a set,and* c_2*,* η_0*, and* η_1 *are positive constants.*

Proof: Proof of the Theorem can be directly obtain as the results of (3.27) and Theorem 2 and Theorem 5.

In the following, we provide the recovery bound for two special cases as an example.

Corollary 7 *If* $R(\delta\theta)$ *is the* L_1 *norm,* $\delta\theta^* \in \mathbb{R}^{p^2}$ *s s-sparse.,* $\Psi(C_r) \leq 4\sqrt{s}$*, and for* $n_2 > cs\log p$*, the recovery error is bounded by*

$$\|\Delta\|_2 \leq c_3\frac{\Psi(C_r)\lambda_{n_1,n_2}}{\kappa} = O\left(\sqrt{\frac{s\log p}{\min(n_1, n_2)}}\right).$$

Corollary 8 *If* $R(\delta\theta)$ *is the group-sparse norm,* $\delta\theta \in \mathbb{R}^{p^2}$*,* $\Psi(C_r) \leq 4\sqrt{s_G}$ *and for* $n_2 \geq c(ms_G + s_G\log N_G)$*, the recovery error is bounded by*

$$\|\Delta\|_2 \leq c_3\frac{\Psi(C_r)\lambda_{n_1,n_2}}{\kappa} = O\left(\sqrt{\frac{s_G m + \log N_G}{\min(n_1, n_2)}}\right).$$

3.4 Experiments

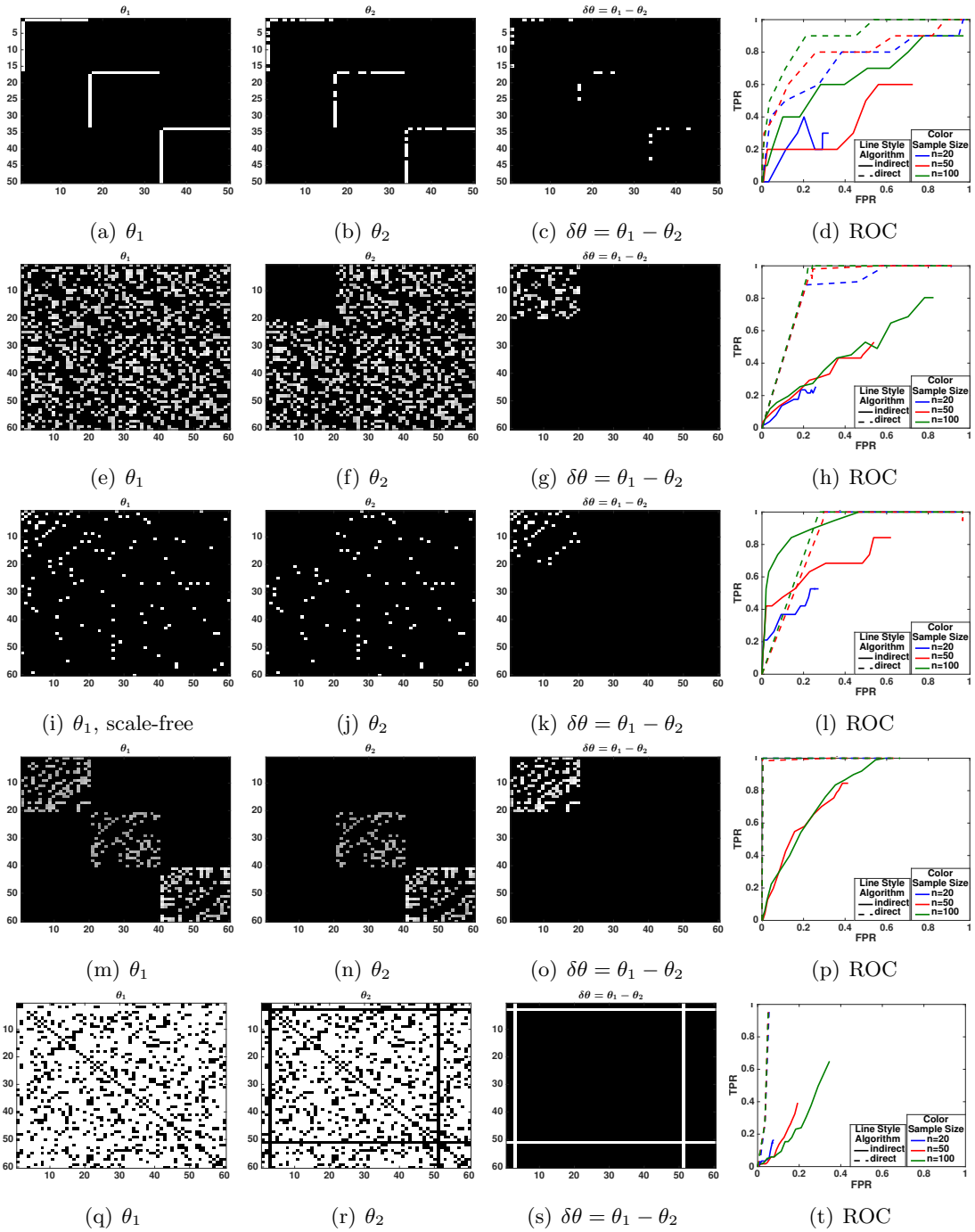
In this Section, we evaluate generalized direct change estimator (direct) with three different norms. and we compare our direct approach with indirect approach. For indirect approach, we first estimate Ising model structures $\hat{\theta}_1$ and $\hat{\theta}_2$ with L_1 norm regularizer, separately [164]. Then, we obtain $\delta\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$. In all experiments, we draw n_1 and n_2 i.i.d samples from each Ising model by running Gibbs sampling. Here we set $n = n_1 = n_2 = \{20, 50, 100\}$.

L_1 **norm:** Here we first generate θ_1^* with three disconnected star sub-graphs (Figure 3.4-a) with $p = 50$. We generate the weights uniformly random between $\{0.3 - 0.5\}$. We then generate θ_2^* by removing 10 random edges from θ_1^* (Figure 3.4-b). It is interesting that although individual graphs are sparse, but direct approach has a better ROC

curve for all values of n (Figure 3.4-d). Similar results obtained by with random graph structure of θ_1^* and θ_2^*.

Group-sparse norm: In this set of experiments, we evaluate direct method with three different structure for θ_1^*: (i) a random graph structure (Figure 3.4-e), (ii) scale free graph structure (Figure 3.4-i), and (iii) a cluster graph structure (Figure 3.4-m). In all settings, we set $p = 60$ and generate θ_2^* by removing a block of edges from θ_1^* (Figure 3.4-(f,j,n)). For random graph structure and block structure, direct method has a better ROC curve (Figure 3.4-h,p). But, for scale-free structure, since the individual graphs are sparse, indirect method can estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ correctly, and thus have a better ROC curve (Figure 3.4-l).

Node perturbation: Here, we first generate a random graph structure θ_1^*, and then generate θ_2^* by perturbing two nodes in θ_1^*. Here we set $p = 60$ and generate θ_2^* by setting rows and columns $3, 51$ to zero in θ_1^* (Figure 3.4-s). Although, the individual graphs are dense but direct approach can estimate edges in $\delta\theta$ with only $n = 20$ samples (Figure 3.4-t).



(a) θ_1 (b) θ_2 (c) $\delta\theta = \theta_1 - \theta_2$ (d) ROC

(e) θ_1 (f) θ_2 (g) $\delta\theta = \theta_1 - \theta_2$ (h) ROC

(i) θ_1, scale-free (j) θ_2 (k) $\delta\theta = \theta_1 - \theta_2$ (l) ROC

(m) θ_1 (n) θ_2 (o) $\delta\theta = \theta_1 - \theta_2$ (p) ROC

(q) θ_1 (r) θ_2 (s) $\delta\theta = \theta_1 - \theta_2$ (t) ROC

Figure 3.1: First row $\delta\theta^*$ has a sparse structure (L_1 norm) and θ_1^* has 3 disconnected star graphs. Second, third, and forth rows $\delta\theta^*$ has group sparse structure (group sparse norm) where θ_1^* has a random graph structure in second row, scale-free structure in third row, and block structure in forth row. Last row $\delta\theta^*$ has two perturbed norm (Node perturbation) and θ_1^* has a random graph structure. Blacks in heatmaps denotes zeros. ROC curve for different structures show in the last column. Direct approach has a better ROC curve for all structures except with scale-free structure of θ_1^*.

Chapter 4

Gaussian Copula Precision Estimation with Missing Values

4.1 Introduction

In this chapter, we propose Double Plug-in Gaussian (DoPinG) copula estimators to deal with missing values, which estimates the sparse precision matrix corresponding to the non-paranormal distribution. DoPingG copula estimators essentially combines two plug-in procedures for dealing with missing values [103] and non-Gaussian data [128], yielding a fairly rich family of estimators to deal with incomplete data from the non-paranormal family. Such estimators consider the following three steps: (1) estimate non-parametric correlations, such as Kendall's tau and Spearman's rho, between all pairs of covariates by suitably disregarding missing values; (2) estimate the non-paranormal correlation matrix using the Kendall's tau or Spearman's rho correlation matrix; (3) plug the estimated correlation matrix into existing sparse precision estimators, e.g., graphical LASSO [15, 76], Dantzig selector [235], CLIME [36], etc.

Our analysis follows the development in [128] with one important difference: the samples we consider can have missing values. We investigate how missing values affect the accuracy of covariance estimation, and in turn precision estimation. In particular, the theoretical analysis of DoPinG copula estimators considers two probability spaces, i.e., probability over samples and probability over missing values. We assume that the data is missing completely at random (MCAR) [103], where any element is missing

with probability δ. We prove that DoPinG copula estimators consistently estimate the non-paranormal correlation matrix at a rate of $O(\frac{1}{(1-\delta)}\sqrt{\frac{\log p}{n}})$.

For estimating the precision matrix, one can use any of the available estimators, such as the graphical lasso [15], graphical Dantzig selector [235], as discussed in [128, 103]. We consider the CLIME estimator [36] for our analysis. The CLIME estimator has strong statistical guarantees for consistency along with rates [36], and also comes with inherent computational advantages [213]. In particular, a large scale distributed algorithm has been developed in [213], which can scale up to millions of dimensions and trillions of parameters, using hundreds of cores. We provide experimental results to show the effect of sample size and percentage of missing data on the model performance. Experimental results show that DoPinG is significantly better than estimators like mGlasso, which are primarily designed for Gaussian data.

The rest of chapter is organized as follows. We propose nonparanormal dual plug-in estimators with missing values in Section 4.2. In Section 4.3, we give the theoretical guarantees in terms of rates of convergences under element-wise L_∞ norm. We present experimental results in Section 4.4.

4.2 Method

We consider a p-dimensional *non-paranormal* distribution [128]. For univariate monotone functions $f_1, \ldots, f_p$ and a positive definite correlation matrix $\Sigma^0 \in \mathbb{R}^{p\times p}$, a p-dimensional random variable $X = (X_1, \ldots, X_p)^T$ has a non-paranormal distribution $X \sim \text{NPN}_p(f, \Sigma^0)$ if $f(X) = (f_1(X_1), \ldots, f_p(X_p)) \sim N_p(0, \Sigma^0)$, a p-dimensional multivariate Gaussian distribution with correlation matrix Σ^0. We focus on estimating the sparse precision matrix $\Omega_0 = \Sigma_0^{-1}$ corresponding to the non-paranormal distribution.

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be samples drawn independently from $\text{NPN}_p(f, \Sigma^0)$. We further assume that for dimension j, x_{ij} will be missing with probability $\delta \in [0, 1]$. Let $b_{ij} = 1$ if x_{ij} is observed, and $b_{ij} = 0$ otherwise. Thus, $P(b_{ij} = 1) = 1 - \delta$. We assume the data is missing completely at random (MCAR) [103].

In order to estimate the precision matrix Ω^0 using CLIME, we need an empirical estimate $\hat{S}_n$ of the correlation matrix Σ^0. In particular, the elementwise L_∞ norm between the matrices need to be suitably bounded for norm consistency of precision

estimation. As shown in [128] , $\hat{S}_n$ can be efficiently computed from the empirical Kendall's tau or Spearman's rho correlation matrix. Hereafter, for ease of notation, we drop the subscript n on $\hat{S}$ and other sample estimates.

DoPinG copula estimators consider three steps in estimating the precision matrix. First, suitably generalizing the plug-in procedure for estimating non-parametric correlations to handle missing values, pairwise Kendall's tau or Spearman's rho correlation between covariates is estimated. Second, the correlation matrix corresponding to the non-paranormal distribution is estimated using the Kendall's tau or Spearman's rho correlation matrices. Third, the precision matrix is estimated by simply plugging in the estimated correlation matrix into existing sparse precision matrix estimators. We discuss each one of these steps below.

4.2.1 Kendall's tau with missing values

Given that samples have missing values, we compute the Kendall's tau for dimensions (j, k) using the n_{jk} *effective* independent samples which have values for both dimensions. In particular, we estimate Kendall's rho as:

$$\hat{\tau}_{jk} = \frac{1}{n_{jk}(n_{jk}-1)} \sum_{\substack{i,i'=1 \\ i \neq i'}}^{n} b_{ij} b_{ik} b_{i'j} b_{i'k} \text{sign}((x_i^j - x_{i'}^j)(x_i^k - x_{i'}^k)) \ , \tag{4.1}$$

where $n_{jk} = \sum_{i=1}^{n} b_{ij} b_{ik}$. Note for the i-th sample, both the j- and k-th dimensions should not be missing. In other words, the samples with missing values will not be considered in the estimation of the Kendall' tau.

The second step is to estimate the correlation matrix directly based on the Kendall's tau. Following [128, 110, 68], we consider the following estimator $\hat{S}^\tau = [\hat{S}^\tau_{jk}]$ for the estimated correlation matrix Σ^0:

$$\hat{S}^\tau_{jk} = \begin{cases} \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if } j = k \ . \end{cases} \tag{4.2}$$

4.2.2 Spearman's rho with missing values

Similar to the estimation of Kendall's tau for missing values, we also compute the Spearman's rho for dimensions (j, k) using the n_{jk} *effective* independent samples which have values for both dimensions. In particular, $n_{jk} = \sum_{i=1}^{n} b_{ij}b_{ik}$. Let r_i^j be the rank of x_i^j among the n_{jk} samples with values and $\bar{r}_{jk}$ be the average, i.e., $\bar{r}_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n} r_i^j b_{ij} b_{ik}$. Spearman's rho is defined as follows:

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^{n}(r_i^j - \bar{r}_{jk})(r_i^k - \bar{r}_{jk})b_{ij}b_{ik}}{\sqrt{\sum_{i=1}^{n}[(r_i^j - \bar{r}_{jk})^2 b_{ij}b_{ik}] \sum_{i=1}^{n}[(r_i^k - \bar{r}_{jk})^2 b_{ij}b_{ik}]}} , \tag{4.3}$$

which is the first step in DoPinG.

Based on the estimate of the Spearman's rho (4.3), following [128, 227] , the second step is to estimate $\hat{S}^{\rho} = [\hat{S}^{\rho}_{jk}]$ for the unknown correlation matrix Σ^0:

$$\hat{S}^{\rho}_{jk} = \begin{cases} 2\sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if } j = k \ . \end{cases} \tag{4.4}$$

4.2.3 Plugin estimate for CLIME

Having obtained $\hat{S}$ ($\hat{S}^{\tau}$ or $\hat{S}^{\rho}$), we can plugin it into any sparse precision estimators, e.g., graphical lasso [15], graphical Dantzig selector [235], CLIME [36]. In particular, we plugin $\hat{S}$ into the CLIME estimator [227]:

$$\hat{\Omega}_n = \text{argmin}_{\hat{\Omega}} \|\hat{\Omega}\|_1 \quad \text{s.t.} \quad \|\hat{S}\hat{\Omega} - \mathbf{I}\|_{\infty} \leq \lambda_n \ , \tag{4.5}$$

where λ_n is a tuning parameter and $\mathbf{I}$ is an identity matrix. The CLIME estimator has strong statistical guarantees [36], and also comes with inherent computational advantages. The estimator can scale up to millions of dimensions and can be run on hundreds of cores [213]. In [213], (4.5) is decomposed into solving $\lceil p/k \rceil$ independent column block linear programs where each column block contains $k (1 \leq k \leq p)$ columns. Denoting $\mathbf{X} \in \Re^{p \times k}$ be k columns of $\hat{\Omega}$, (4.5) can be written as

$$\min \|\mathbf{P}\|_1 \quad \text{s.t.} \quad \|\hat{S}\mathbf{P} - \mathbf{E}\|_{\infty} \leq \lambda_n \ , \tag{4.6}$$

which can be solved by an inexact ADMM algorithm [27, 212] given in Algorithm 2 [213] where ρ, η are parameters of ADMM and

$$soft(\mathbf{P}, \gamma) = \begin{cases} P_{ij} - \gamma \ , & \text{if } \ P_{ij} > \gamma \ , \\ P_{ij} + \gamma \ , & \text{if } \ P_{ij} < -\gamma \ , \\ 0 \ , & \text{otherwise} \end{cases}$$

$$box(\mathbf{P}, \mathbf{E}, \lambda_n) = \begin{cases} E_{ij} + \lambda \ , & \text{if } \ P_{ij} - E_{ij} > \lambda_n \ , \\ P_{ij} \ , & \text{if } \ |P_{ij} - E_{ij}| \leq \lambda_n \ , \\ E_{ij} - \lambda \ , & \text{if } \ P_{ij} - E_{ij} < -\lambda_n \ , \end{cases}$$

While steps 5, 7, 8 and 10 amount to elementwise operations, the most intensive computation is matrix multiplication in steps 6 and 9 which can be solved in parallel.

Note that the estimated correlation matrix $\hat{S}$ ($\hat{S}^\tau$ or $\hat{S}^\rho$) may be not positive semi-definite. Sparse precision estimators do require the positive semi-definiteness assumption in theory and most algorithms may fail if the input correlation matrix is not positive semi-definite [128, 103]. The inexact ADMM algorithm for CLIME in Algorithm 2 does not necessarily require $\hat{S}$ to be positive semi-definite. As long as the linear programs (4.5) have solutions, Algorithm 2 still works, although there is no guarantee that the solution is positive definite. Therefore, one may project the input correlation matrix onto the cone of positive semi-definite matrix in order to obtain a positive definite precision matrix with high probability using Algorithm 2. We study the effect of the two choices on the performance of DoPinG in experiments in Section 4.

4.3 Theoretical Analysis

In this section, we present statistical guarantees for the proposed DoPinG by leveraging existing analysis in [128, 36, 227]. Note that the consistency analysis of the CLIME estimate $\hat{\Omega}$ relies on obtaining a consistent estimate of the covariance Σ^0, defined in terms of the elementwise L_∞ norm of the difference $(\hat{S} - \Sigma^0)$. Therefore, we first analyze $\sup_{jk} \left|\hat{S}^\tau_{jk} - \Sigma^0_{jk}\right|$ for the Kendall's tau ($\hat{S} = \hat{S}^\tau$) and Spearman's rho ($\hat{S} = \hat{S}^\rho$) separately. Our proof operates on two probability spaces, i.e., probabilities over the samples $\mathbb{P}_X$ and probabilities over the Bernoulli missing values $\mathbb{P}_B$. Then, we plug the results into the consistency analysis of the CLIME to obtain the optimal statistical rate

Algorithm 2 Column Block Inexact ADMM for CLIME

Input: $\hat{S}, \lambda_n, \rho, \eta$
Output: $\mathbf{P}$
Initialization: $\mathbf{P}^0, \mathbf{Z}^0, \mathbf{Y}^0, \mathbf{V}^0, \hat{\mathbf{V}}^0 = 0$
for $t = 0$ to $T-1$ **do**
  X-update: $\mathbf{P}^{t+1} = soft(\mathbf{P}^t - \mathbf{V}^t, \frac{1}{\eta})$, where
  Mat-Mul: $\mathbf{U}^{t+1} = \hat{S}\mathbf{P}^{t+1}$
  Z-update: $\mathbf{Z}^{t+1} = box(\mathbf{U}^{t+1} + \mathbf{Y}^t, \lambda_n)$, where
  Y-update: $\mathbf{Y}^{t+1} = \mathbf{Y}^t + \mathbf{U}^{t+1} - \mathbf{Z}^{t+1}$
  Mat-Mul: $\hat{\mathbf{V}}^{t+1} = \hat{S}\mathbf{Y}^{t+1}$
  V-update: $\mathbf{V}^{t+1} = \frac{\rho}{\eta}(2\hat{\mathbf{V}}^{t+1} - \hat{\mathbf{V}}^t)$
end for

of convergence.

We first consider the probabilities over missing values in the following Lemma which we need in the analysis of Kendall's tau and Spearman's rho:

Lemma 9 *Let $B = [b_{ij}] \in \{0,1\}^{n\times p}$ be an binary matrix. Assume b_{ij} is i.i.d. with a Bernoulli distribution where $P(b_{ij} = 0) = \delta$ and $P(b_{ij} = 1) = 1 - \delta$. Let $n_{jk} = \sum_{i=1}^n b_{ij}b_{ik}$. For any $m > 0$, and any $0 < \epsilon < 1$, we have*

$$\mathbb{P}_B\left(\sum_{j,k}\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right) > \frac{1}{p^m}\right\}$$
$$\leq \exp\left(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p)\right) \ , \tag{4.7}$$

Proof: Since n_{jk} is a sum of n independent Bernoulli random variables $b_{ij}b_{ik}$ with $P(b_{ij}b_{ik} = 1) = (1-\delta)^2$, by linearity of expectation and independence of samples, we have $E[n_{jk}] = \sum_{i=1}^n E[b_{ij}b_{ik}] = n(1-\delta)^2$. By standard Chernoff bounds, for any $\epsilon < 1$, we have

$$\mathbb{P}_B\left(n_{jk} < E[n_{jk}](1-\epsilon)\right) \leq \exp\left(-\epsilon^2(1-\delta)^2 n/2\right)$$
$$\Rightarrow \mathbb{P}_B\left(\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\} \geq \frac{1}{p^{m+2}}\right)$$
$$\leq \exp\left(-\epsilon^2(1-\delta)^2 n/2\right) \ , \tag{4.8}$$

where we have substituted the expectation $E[n_{jk}]$. By considering probabilities over the missing values, we have

$$\begin{aligned}
&\mathbb{P}_B\left(\sum_{j,k}\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\}>\frac{1}{p^m}\right)\\
&\leq \sum_{j,k}\mathbb{P}_B\left(\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\}>\frac{1}{p^{m+2}}\right)\\
&\leq p^2\exp\left(-\epsilon^2(1-\delta)^2n/2\right)\\
&= \exp\left(-(\epsilon^2(1-\delta)^2n/2-2\log p)\right)\ , \qquad (4.9)
\end{aligned}$$

which completes the proof. ■

4.3.1 Kendall's Tau with Missing Values

The following theorem shows that $\sup_{jk}\ \left|\hat{S}^{\tau}_{jk}-\Sigma^0_{jk}\right| \leq O(\sqrt{\log p/n})$ with high probability.

Theorem 10 *For any $n \geq 1$, for any $m > 0$, and any $0 < \epsilon < 1$, with probability at least $(1-\frac{1}{p^m})(1-\exp(-(\epsilon^2(1-\delta)^2n/2-2\log p))$, we have*

$$\sup_{jk}\ \left|\hat{S}^{\tau}_{jk}-\Sigma^0_{jk}\right| \leq \frac{\pi}{1-\delta}\ \sqrt{\frac{m+2}{1-\epsilon}}\ \sqrt{\frac{\log p}{n}}\ . \qquad (4.10)$$

Proof: Since $\hat{\tau}_{jk}$ is an unbiased estimator of τ_{jk}, $E[\hat{\tau}_{jk}]=\tau_{jk}$. Using (4.2), we have

$$\begin{aligned}
&\mathbb{P}_X\left(\left|\hat{S}_{jk}-\Sigma^0_{jk}\right|>t\right)\\
&=\mathbb{P}_X\left(\left|\sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right)-\sin\left(\frac{\pi}{2}\tau_{jk}\right)\right|>t\right)\\
&\leq \mathbb{P}_X\left(|\hat{\tau}_{jk}-\tau_{jk}|>\frac{2}{\pi}t\right)\\
&\leq \exp\left(-\frac{n_{jk}t^2}{\pi^2}\right)\ , \qquad (4.11)
\end{aligned}$$

where the last inequality uses the Hoeffding bound for the U-statistics [128, 91]. Application of the union bound yields

$$\begin{aligned}&\mathbb{P}_X\left(\sup_{jk}\left|\hat{S}^{\tau}_{jk}-\Sigma^0_{jk}\right|>t\right)\\&\leq\sum_{j,k}\exp\left(-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right)\ ,\end{aligned}\tag{4.12}$$

where we have substituted $t=\frac{\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$. The bound in the above form is itself a random variable, and the elements of the sum are identically distributed but are not independent.

By considering probabilities over the missing values and using Lemma 9, we have

$$\begin{aligned}&\mathbb{P}_B\left(\mathbb{P}_X\left(\sup_{jk}\left|\hat{S}^{\tau}_{jk}-\Sigma^0_{jk}\right|\leq t\right)\geq\left(1-\frac{1}{p^m}\right)\right)\\&\geq 1-\exp\left(-(\epsilon^2(1-\delta)^2n/2-2\log p)\right)\ .\end{aligned}\tag{4.13}$$

Noting that the random variables (X,B) are independent completes the proof. ∎

4.3.2 Spearman's Rho with Missing Values

As we work on the n_{jk} effective samples wth values by disregarding missing values, we can leverage the analysis in [128] except n_{jk} is a random variable. Following [128], (4.3) can be rewritten as [90, 128]:

$$\begin{aligned}\hat{\rho}_{jk}&=\frac{3\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\text{sign}(x_i^j-x_s^j)(x_i^k-x_t^k)b_{ij}b_{ik}b_{sj}b_{sk}b_{tj}b_{tk}}{n_{jk}^3-n_{jk}}\\&=\frac{n_{jk}-2}{n_{jk}+1}U_{jk}+\frac{3}{n_{jk}+1}\hat{\tau}_{jk}\ .\end{aligned}\tag{4.14}$$

where $\hat{\tau}_{jk}$ is Kendall's tau statistics and U_{jk} is a 3rd-order U-statistics

$$U_{jk}=\frac{3\sum_{i\neq s\neq t}\text{sign}(x_i^j-x_s^j)(x_i^k-x_t^k)b_{ij}b_{ik}b_{sj}b_{sk}b_{tj}b_{tk}}{n_{jk}(n_{jk}-1)(n_{jk}-2)}\ .\tag{4.15}$$

Note $n_{jk} = \sum_{i=1}^{n} b_{ij} b_{ik}$ is a sum of n independent Bernoulli random variables $b_{ij} b_{ik}$ with $\mathbb{E}(n_{ij}) = (1-\delta)^2 n$.

Theorem 11 *For any $m > 0$, $0 < \epsilon < 1$, and*

$$n \geq \frac{36}{(m+2)(1-\epsilon)(1-\delta)^2 \log p} \ , \tag{4.16}$$

with probability at least $(1 - \frac{1}{p^m})(1 - \exp(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p))$, we have

$$\sup_{jk} \ \left| \hat{S}^{\tau}_{jk} - \Sigma^{0}_{jk} \right| \leq \frac{4\pi}{1-\delta} \sqrt{\frac{m+2}{1-\epsilon}} \sqrt{\frac{\log p}{n}} \ . \tag{4.17}$$

Proof: Let $0 < \alpha < 1$. According to (4.14), we have

$$\begin{aligned} \mathbb{P}_X(|\hat{\rho}_{jk} - \mathbb{E}(\hat{\rho}_{jk})| > t) &\leq \mathbb{P}_X(|U_{jk} - \mathbb{E}(U_{jk})| > \alpha t) \\ &+ \mathbb{P}_X\left(\frac{3}{n_{jk}+1} |\hat{\tau}_{jk} - \tau_{jk}| > (1-\alpha)t \right) \ . \end{aligned} \tag{4.18}$$

Since $-1 \leq \tau_{jk} \leq 1$, $|\hat{\tau}_{jk} - \tau_{jk}| \leq 2$, then

$$\begin{aligned} &\mathbb{P}_X\left(\frac{3}{n_{jk}+1} |\hat{\tau}_{jk} - \tau_{jk}| > (1-\alpha)t \right) \\ &\leq \mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t \right) \ . \end{aligned} \tag{4.19}$$

Applying Hoeffding's bound for U-statistics, we have

$$\begin{aligned} &\mathbb{P}_X(|U_{jk} - \mathbb{E}(U_{jk})| > \alpha t) \\ &\leq \exp\left(-2 \left\lfloor \frac{n_{jk}}{3} \right\rfloor \frac{\alpha^2 t^2}{36} \right) = \exp\left(-\frac{n_{jk} \alpha^2 t^2}{54} \right) \ . \end{aligned} \tag{4.20}$$

Combining (4.19) and (4.20) yields

$$\begin{aligned} \mathbb{P}_X\left(|\hat{\rho}_{jk} - \ \mathbb{E}(\hat{\rho}_{jk})| > t \right) &\leq \exp\left(-\frac{n_{jk} \alpha^2 t^2}{54} \right) \\ &+ \mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t \right) \ . \end{aligned} \tag{4.21}$$

In particular, if $n_{jk} \geq \frac{6}{(1-\alpha)t}$, the second term on the RHS is 0. Since $\hat{\rho}_{jk}$ is a biased estimator, following [128], we use the following bias equation [242]:

$$\mathbb{E}\hat{\rho}_{jk} = \frac{6}{\pi(n_{jk}+1)}\left[\arcsin(\Sigma^0_{jk}) + (n_{jk}-2)\arcsin(\frac{\Sigma^0_{jk}}{2})\right] \ . \tag{4.22}$$

Note we only use n_{jk} effective number of samples. Thus,

$$\Sigma^0_{jk} = 2\sin\left(\frac{\pi}{2}\mathbb{E}\hat{\rho}_{jk} + a_{jk}\right) \ , \tag{4.23}$$

where

$$a_{jk} = \frac{\pi\mathbb{E}\hat{\rho}_{jk} - 2\arcsin(\Sigma^0_{jk})}{2(n_{jk}-2)} \ , |a_{jk}| \leq \frac{\pi}{n_{jk}-2} \ . \tag{4.24}$$

If $n_{jk} \geq \frac{6\pi}{t} + 2$, $|a_{jk}| \leq \frac{t}{6}$. Therefore, the analysis is simplified if $\inf_{jk} n_{jk} \geq c_0$ where

$$c_0 \geq \max\left\{\frac{6}{(1-\alpha)t}, \frac{6\pi}{t} + 2\right\} \ . \tag{4.25}$$

Setting $\alpha = \frac{3\sqrt{6}}{8}, t = \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$, we have

$$\begin{aligned}
\frac{6}{(1-\alpha)t} &\leq \frac{24\pi}{t} = 6(1-\delta)\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}} \ , \\
\frac{6\pi}{t} + 2 &= \frac{3(1-\delta)}{2}\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}} \ .
\end{aligned}$$

Therefore, we choose

$$c_0 = 6(1-\delta)\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}} \ . \tag{4.26}$$

Define an event $Z = \{\inf_{jk} n_{jk} \geq c_0\}$, and let $\bar{Z}$ be the complement of the event. Further, the event of interest is $Y = \left\{\sup_{j,k}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}\right\}$. Then, the probability of the event of interest can be lower bounded as:

$$\begin{aligned}
P(Y) &= P(Y|Z)P(Z) + P(Y|\bar{Z})P(\bar{Z}) \\
&\geq P(Y|Z)P(Z) \ .
\end{aligned} \tag{4.27}$$

Next, we focus on getting lower bounds to both $P(Z)$ and $P(Y|Z)$.

Note $n_{jk} = \sum_{i=1}^{n} b_{ij} b_{ik}$ and $\mathbb{E}[n_{jk}] = (1-\delta)^2 n$, using Chernoff bounds,

$$\mathbb{P}_B\left(n_{jk} < (1-\epsilon)(1-\delta)^2 n\right) \leq \exp\left(-\epsilon^2 (1-\delta)^2 n/2\right) \ . \tag{4.28}$$

By the union bound,

$$\begin{aligned} &\mathbb{P}_B\left(\inf_{jk} n_{jk} < (1-\epsilon)(1-\delta)^2 n\right) \\ &\leq \exp\left(-\epsilon^2 (1-\delta)^2 n/2 + 2\log p\right) \ , \end{aligned} \tag{4.29}$$

which is equivalent to

$$\begin{aligned} &\mathbb{P}_B\left(\inf_{jk} n_{jk} \geq (1-\epsilon)(1-\delta)^2 n\right) \\ &\geq 1 - \exp\left(-\epsilon^2 (1-\delta)^2 n/2 + 2\log p\right) \ . \end{aligned} \tag{4.30}$$

If $(1-\epsilon)(1-\delta)^2 n \geq c_0$, i.e.,

$$n \geq \frac{36}{(m+2)(1-\epsilon)(1-\delta)^2 \log p} \ , \tag{4.31}$$

then

$$\mathbb{P}_B\left(\inf_{jk} n_{jk} \geq c_0\right) \geq 1 - \exp\left(-\epsilon^2 (1-\delta)^2 n/2 + 2\log p\right) \ , \tag{4.32}$$

which gives a lower bound to $P(Z)$ as desired. Now, conditioned on Z, i.e., $\inf_{jk} n_{jk} \geq c_0$, we have $|a_{jk}| \leq \frac{t}{6}$, and $\mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t \,\middle|\, Z\right) = 0$. Assuming n satisfies (4.31) and using (4.21), (4.23), we have

$$\begin{aligned} &\mathbb{P}_X\left(|\hat{S}^{\rho}_{jk} - \Sigma^0_{jk}| > t \,\middle|\, Z\right) \\ &= \mathbb{P}_X\left(\left|2\sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right) - 2\sin(\frac{\pi}{6}\mathbb{E}\hat{\rho}_{jk} + a_{jk})\right| > t \,\middle|\, Z\right) \\ &\leq \mathbb{P}_X\left(\left|\frac{\pi}{3}\hat{\rho}_{jk} - \frac{\pi}{3}\mathbb{E}\hat{\rho}_{jk} - 2a_{jk}\right| > t \,\middle|\, Z\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}_X\left(\left|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk} - \frac{6}{\pi}a_{jk}\right| > \frac{3t}{\pi}\ \middle|Z\right) \\
&\leq \mathbb{P}_X\left(|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk}| > \frac{3t}{\pi} - \left|\frac{6}{\pi}a_{jk}\right|\ \middle|Z\right) \\
&\leq \mathbb{P}_X\left(|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk}| > \frac{2t}{\pi}\ \middle|Z\right) \\
&\leq \exp\left(-\frac{2n_{jk}\alpha^2 t^2}{27\pi^2}\right)\ , \qquad (4.33)
\end{aligned}
$$

where the conditioning on Z, i.e., $\{\inf_{j,k} n_{jk} \geq c_0\}$, has been dropped in the last inequality yielding an upper bound. Setting $\alpha = \frac{3\sqrt{6}}{8}, t = \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$, by the union bound, we have

$$
\begin{aligned}
&\mathbb{P}_X\left(\sup_{jk}|\hat{S}^{\rho}_{jk} - \Sigma^0_{jk}| > t\ \middle|Z\right) \\
&\leq \sum_{j,k}\exp\left(-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right)\ , \qquad (4.34)
\end{aligned}
$$

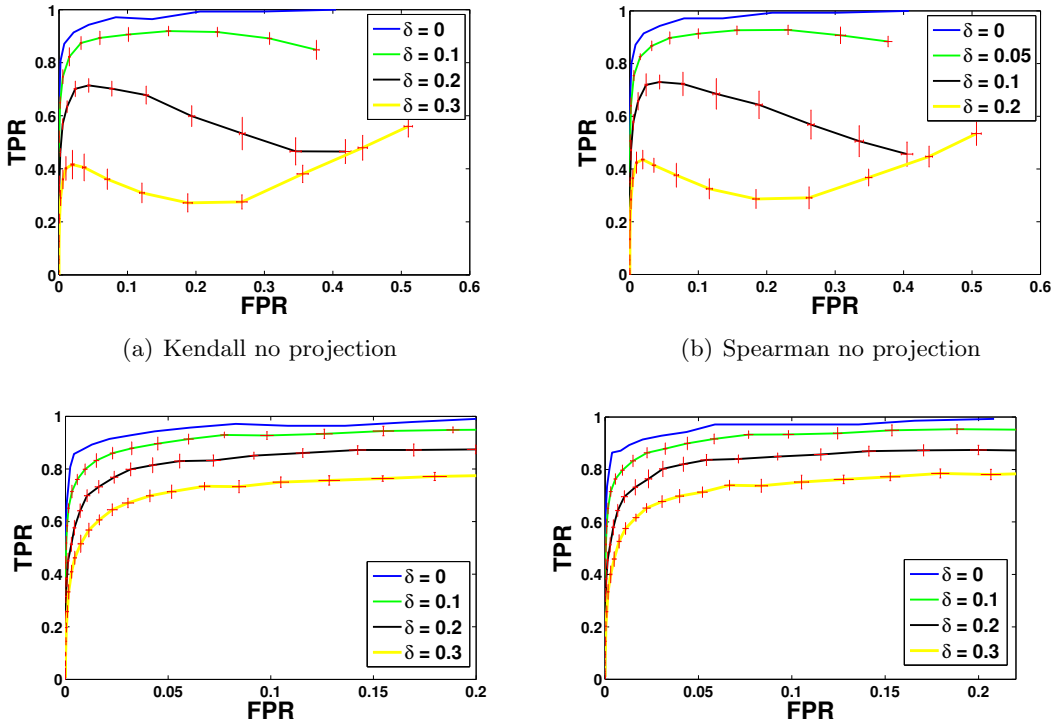
which is the same as (4.12). Using Lemma 9, we then have $P(Y|Z) \geq \left(1 - \frac{1}{p^m}\right)$. The result of the theorem then follows from (4.27) and (4.30). ∎

4.3.3 Plug-in CLIME Estimator

Since $\hat{S}$ ($\hat{S}^{\tau}$ or $\hat{S}^{\rho}$) satisfies (4.10) or (4.17) with high probability, choosing $\lambda_n \geq \frac{\pi\|\Omega^0\|_{L_1}}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$ or $\lambda_n \geq \frac{4\pi\|\Omega^0\|_{L_1}}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$ ensures that the conditions for consistency of the CLIME estimate $\hat{\Omega}$ are satisfied. The CLIME estimator considers the following family of precision matrices $\mathcal{U} = \mathcal{U}(M, q, s_0(p)) = \Big\{\Omega : \Omega \succ 0, \|\Omega\|_{L_1} \leq M, \max_{1\leq i\leq p}\sum_{j=1}^{p}|\omega_{ij}|^q \leq s_0(p)\Big\}$, for $0 \leq q < 1$. Then, the CLIME estimator has the following guarantees:

Theorem 12 *Let $\Omega_0 \in \mathcal{U}(M, q, s_0(p))$. If $\lambda_n \geq \|\Omega_0\|_{L_1}\max_{ij}|\hat{\sigma}_{n,ij} - \sigma_{0,ij}|$, then we have*

$$
\begin{aligned}
|\hat{\Omega}_n - \Omega_0|_{\infty} &\leq 4\|\Omega_0\|_{L_1}\lambda_n\ , \qquad (4.35) \\
\|\hat{\Omega}_n - \Omega_0\|_2 &\leq Cs_0(p)(4\|\Omega_0\|_{L_1})^{1-q}\lambda_n^{1-q}\ , \qquad (4.36)
\end{aligned}
$$



(a) Kendall no projection

(b) Spearman no projection

(c) Kendall, projection

(d) Spearman, projection

Figure 4.1: (a,b) ROC curves without projection ($\hat{S}$ need not be positive semi-definite), (c,d) ROC curves with projection ($\hat{S}$ is positive semi-definite) with $n = 200$ and under different missing probabilities ($\delta = 0.1 - 0.3$). By increasing number of observed data (smaller δ), the ROC curve approaches the ROC curve of no-missing data ($\delta = 0$).

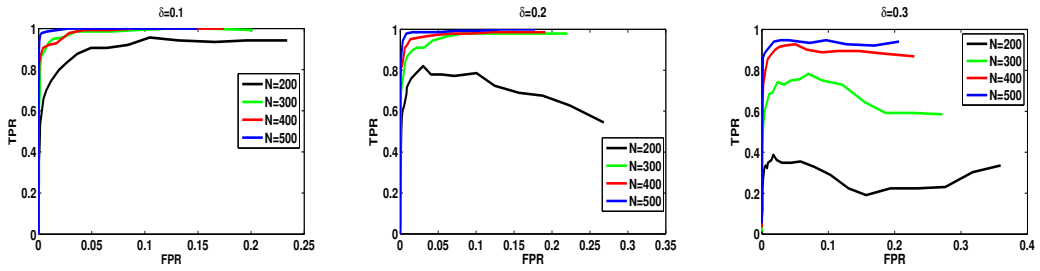
$$\frac{1}{p}\|\hat{\Omega}_n - \Omega_0\|_F^2 \leq Cs_0(p)(4\|\Omega_0\|_{L_1})^{2-q}\lambda_n^{2-q} \ , \tag{4.37}$$

where $C \leq 2(1 + 2^{1-q} + 3^{1-q})$ *is a constant.*

Note that deterministic bounds in Theorem 12 for precision estimation relies on $|\hat{\Sigma}_n - \Sigma_0|_\infty = \max_{i,j} |\hat{\sigma}_{n,ij} - \sigma_{0,ij}|$.

4.4 Experimental Results

We present experimental results of DoPinG on both synthetic datasets and real datasets to illustrate model performance. The first set of experiments on synthetic data illustrate



(a) $\delta = 10\%$ (b) $\delta = 20\%$ (c) $\delta = 30\%$

Figure 4.2: ROC curve with $\delta = 0.1, 0.2, 0.3$, $p = 100$, and different number of samples (n). For a fixed value of δ, with increasing number of samples, the higher TP rates is obtained.

the effect of sample size and percentage of missing data on model performance. Then we compare DoPinG with mGlasso on both synthetic data and climate dataset.

4.4.1 Synthetic Data

To generate synthetic data, we use the procedure described in [128]. First, a d-dimensional sparse graph $G = (V, E)$ is generated as follows: Let $V = \{1, ..., p\}$ correspond to variables $X = (X_1, ..., X_d)$. We associate each index j with a bivariate point $Y_j = (Y_j^{(1)}, Y_j^{(2)}) \in [0, 1]^2$ where each $Y_j^{(k)} \sim \text{Unif}[0, 1]$, $k = 1, 2$, $j \in \{1, \cdots, d\}$. An edge is associated between vertices (i, j) with probability of $P((i, j) \in E) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|y_i - y_j\|^2}{0.25}\right)$ where $y_j = (y_j^{(1)}, y_j^{(2)})$ is the observation of Y_j and $\| \cdot \|$ denotes the Euclidean distance. The maximum degree of the graph is limited to 4. Thereafter, n samples are drawn from $NPN_d(f^0, \Sigma^0)$ where f^0 is the Gaussian CDF Transformation with mean 0.05 and standard deviation 0.4. Here, we choose $n = 200$, $p = 100$, and $\delta \in \{0.1, 0.2, 0.3\}$. The final results shown below are averages over 10 experimental runs for both Kendall's tau and Spearman's rho. The ROC curve is generated by varying the tuning parameter λ in the CLIME and calculating the corresponding False Positive Rate (FPR) and True Positive Rate (TPR) [128].

First, we directly run Algorithm 2 using $\hat{S}$ ($\hat{S}^\tau$ or $\hat{S}^\rho$) estimated using Kendall's tau and Spearman's rho. The ROC curve with different probabilities of missing values is plotted in Figure 4.1. We observe that the performance of Kendall's tau and Spearman's rho is almost the same for the same percentage of missing values. Note that the tuning

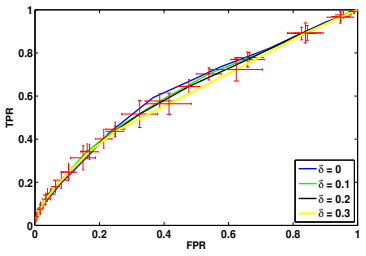
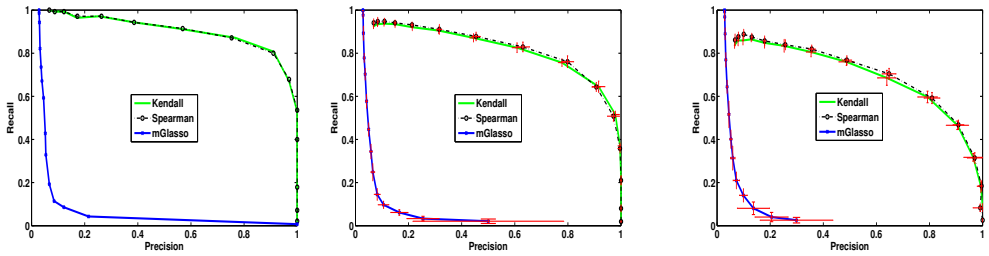


Figure 4.3: ROC curve of mGlasso with $n = 200$ and different missing probabilities. mGlasso has a worse performance on non-Gaussian data compared to DoPinG (Figure 4.1).

parameter λ controls the sparsity of the estimated graph, i.e., a small value of λ provides a dense graph. When λ is large enough the predicted edges are all among the correct edges leading to a zero FPR. By decreasing λ, false edges that are not in the original graph are added, i.e., increasing FPR and saturating TPR. It shows that the estimator is conservative in adding edges. Figure 4.1 also illustrates that increasing number of missing values (increasing δ) deteriorates model performance, while increasing variance of estimate.

As mentioned in section 4.2.3, the estimated correlation matrix $\hat{S}$ may be not positive semi-definite. Therefore, we project $\hat{S}$ into the positive semi-definite (PSD) cone, and execute Algorithm 2 using the PSD matrix. Figures 4.1 (c,d) plot the ROC curve with projection for Kendall's tau and Spearman's rho respectively. For small δ, e.g. $\delta = 0.1$, to some degree, the performances with and without projection are similar. However, when more values are missing, PSD projection greatly improves performance. Increasing percentage of missing values lead to more and larger negative eigenvalues in $\hat{S}$, and performance worsens for higher δ. Note that our analysis shows that the *effective* sample size is $(1-\delta)^2 n$, and decrease of the recovery rate (TPR) with decreasing *effective* sample size is in accordance with our analysis. In other words, for a fixed n the *effective* sample size is smaller for a larger value of δ and therefore, DoPinG has a worse performance with larger value of δ.

Figure 4.2 shows the effect of sample size n with different value of δ on the performance without projection. Under higher percentage of missing values (Figure 4.2(c)), the performance of the method suffers much more with low sample size, compared to data with lower percentage of missing entries (Figure 4.2(a)). In particular, with a



(a) $\delta = 0$ (b) $\delta = 10\%$ (c) $\delta = 20\%$

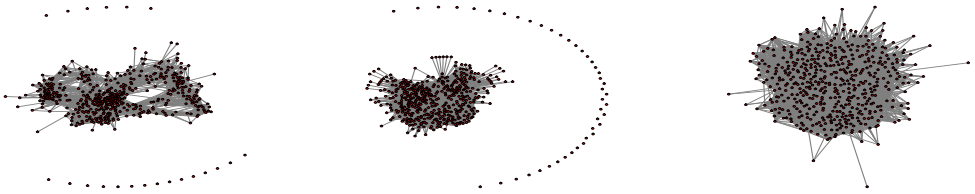
Figure 4.4: Precision and Recall Curve with different δ. DoPinG is significantly better than mGlasso for non-Gaussian data.

sample size $n = 200$ and 30% of missing data, the *effective* sample size is ~ 100 while with 10% of missing data, the *effective* sample size is ~ 160. As a result, to achieve similar recovery rates (TPR,FPR), higher sample size is needed when more percentage of the data is missing.

We compare DoPinG with mGlasso [103] on the synthetic data. The ROC curve of mGlasso is plotted in Figure 4.3. Since mGlasso is designed primarily for Gaussian data, Figure 4.3 clearly illustrates that mGlasso is not suitable for non-Gaussian data. We also plot the precision and recall curve with different probabilities of missing values ($\delta = 0, 0.1, 0.2$) in Figure 4.4. The performance of DoPinG is significantly better than mGlasso.

4.4.2 Climate Data

We compare DoPinG (Spearman's rho) and mGlasso on Climate data. The climate dataset that we use is obtained from the CMIP5 archive, where we use the temperature predicted over land locations by a climate model. We reduce the resolution of the data, since we use it only for illustrative purposes, so that the data contains 500 locations (dimensionality), and yearly averaged samples over 100 years (sample size =100). We randomly remove $\delta = 20\%$ of the entries. We try different λ and report the results which have similar number of edges. In particular, we pick the graph with 12740 edges for DoPinG ($\lambda = 0.02$) as illustrated in Figure 4.5(a). We pick two graphs for mGlasso. One has 8778 edges ($\lambda = 0.001$) and the other has 11860 edges ($\lambda = 0.002$), as shown in Figure 4.5(b) and 4.5(c) respectively. It seems that DoPinG discovers some interesting



(a) DoPinG (12240 edges) (b) mGlasso (8778 edges) (c) mGlasso (11860 edges)

Figure 4.5: The graph discovered by DoPinG and mGlasso.

Table 4.1: Edges dicovered by DoPinG and mGlasso on Climate Data. $>$ denotes the number of edges in DoPinG graph but not in mGlasso graph. $<$ is on the contrary.

Edge No.		Edge Diff	
DoPinG	mGlasso	$>$	$<$
12240	8778	7942	4480
12240	11860	7534	7154

sparsity patterns while mGlasso graphs are messy. In Table 1, we present the difference between DoPinG graph and mGlasso graph. With similar total number of edges, DoPinG graph shows more structure than mGlasso graph. We plan to further investigate this behavior in future work.

Part II

Low Rank Matrix Completion

Chapter 5

Collapsed Monte Carlo Inference for Matrix Completion

5.1 Introduction

In this chapter, we first illustrate that the $\mathcal{MGIG}$ distribution is unimodal where the mode can be obtained by solving an *Algebraic Riccati Equation (ARE) [28].* This characterization leads to an effective importance sampler for the $\mathcal{MGIG}$ distribution. More specifically, for estimating the expectation $\mathbb{E}_{X\sim\mathcal{MGIG}}[g(X)]$, we select a proposal distribution over space of symmetric positive definite matrices like Wishart or Inverse Wishart distribution such that the mode of the proposal matches the mode of the $\mathcal{MGIG}$. As a result, unlike the current sampler [229, 233], by aligning the shape of the proposal and the $\mathcal{MGIG}$, the density of the proposal gets higher values in the high density regions of the target, yielding to a good approximation of $\mathbb{E}_{X\sim\mathcal{MGIG}}[g(X)]$.

Further, we discuss a new application of the $\mathcal{MGIG}$ distribution in latent factor models such as probabilistic matrix factorization (PMF) [174] or Bayesian PCA (BPCA) [21] . In these settings, the given matrix $X \in \mathbb{R}^{N\times M}$ is approximated by a low-rank matrix $\hat{X} = UV^T$ where $U \in \mathbb{R}^{N\times D}$ and $V \in \mathbb{R}^{M\times D}$ with Gaussian priors over the latent matrices U and V. We show that after analytically marginalizing one of the latent matrices in PMF (or BPCA), the posterior over the other matrix has the $\mathcal{MGIG}$ distribution. This illustration yields to a novel Collapsed Monte Carlo (CMC) inference algorithm for PMF. In particular, we marginalize one of the latent matrices, say V,

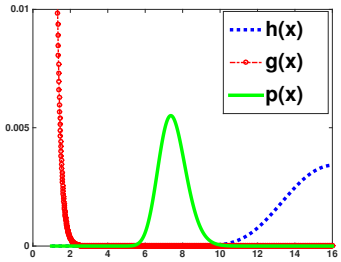


Figure 5.1: An illustration of bad proposal distribution in importance sampling. Let $p(x) = h^*(x)g^*(x)/Z_p \propto h(x)g(x)$. Neither $h(x) = h^*(x)/Z_h$ nor $g(x) = g^*(x)/Z_g$ are a good candidate proposal distribution since their modes are far away from the one of $p(x)$.

and propose a direct Monte Carlo sampling from the posterior of the other matrix, say U. Through extensive experimental analysis on synthetic, SNP, gene expression, and MovieLens datasets, we show that CMC has lower log loss or perplexity with fewer samples than Markov Chain Monte Carlo (MCMC) inference approach for PMF [175].

The rest of the chapter is organized as follows. In Section 5.2, we cover background materials. In Section 5.3, we show that $\mathcal{MGIG}$ is unimodal and give a novel importance sampler for $\mathcal{MGIG}$. We provide the connection of $\mathcal{MGIG}$ with PMF in Section 5.4, present the results in Section 5.5.

5.2 Background and Preliminary

In this section, we provide some background on the relevant topics and tools that will be used in our analysis. We start by an introduction to importance sampling, $\mathcal{MGIG}$ distribution, followed by a brief overview of the ARE.

Notations: Let $\mathbb{S}_{++}^N$ and $\mathbb{S}_{+}^N$ denote the space of symmetric $(N \times N)$ positive definite and positive semi-definite matrix, respectively. Let $|.|$ denote the determinant of matrix, Tr(.) be the matrix trace. A matrix $\Lambda \in \mathbb{S}_{++}^N$ has a Wishart distribution denoted as $\mathcal{W}_N(\Lambda|\Phi, \tau)$ where $\tau > N - 1$ and $\Phi \in \mathbb{S}_{++}^N$ [222]. A matrix $\Lambda \in \mathbb{S}_{++}^N$ has an Inverse Wishart distribution denoted as $\mathcal{IW}_N(\Lambda|\Psi, \alpha)$ where $\alpha > N - 1$ and $\Psi \in \mathbb{S}_{++}^N$ is the scale matrix. We denote $\mathbf{x}_{:m}$ as the m^{th} column of matrix $X \in \mathbb{R}^{N \times M}$ and $\mathbf{x}_n$ as the n^{th} row of X.

5.2.1 Importance Sampling

Consider distribution $p(x) = \frac{1}{Z_p}p^*(x)$ where Z_p is the partition function which plays the role of a normalizing constant. Importance sampling is a general technique for estimating $\mathbb{E}_{x\sim p(x)}[g(x)]$ where sampling from $p(x)$ (the target distribution) is difficult but we can evaluate the value of $p^*(x)$ at any given x [137]. The idea is to draw S samples $\{x_i\}_{i=1}^S$ from a similar but easier distribution denoted by proposal distribution $q(x) = \frac{1}{Z_q}q^*(x)$. Define $w(x_i) = \frac{p^*(x_i)}{q^*(x_i)}$ as the weight of each sample i. Then, we calculate the expected value as follows

$$\mathbb{E}_{x\sim p}[g(x)] = \mathbb{E}_{x\sim q}\left[\frac{g(x)p(x)}{q(x)}\right] \approx \frac{\sum_{i=1}^S g(x_i)w(x_i)}{\sum_{i=1}^S w(x_i)},$$

The efficiency of importance sampling depends on how closely the proposal approximates the target in the shape. One way for monitoring the efficiency of importance sampling is the effective sample size $ESS = \frac{(\sum_{i=1}^S w(x_i))^2}{\sum_{i=1}^S w^2(x_i)}$ [105]. Very small value of ESS indicates a big discrepancy between the proposal and target (for example when the mode of the proposal distribution is far away from the target's mode) leading to a drastically wrong estimate of $\mathbb{E}_{x\sim p}[g(x)]$ [137].

5.2.2 MGIG Distribution

$\mathcal{MGIG}$ distribution was first introduced in [19] as a distribution over the space of symmetric $(N \times N)$ positive definite matrices defined as follows.

Definition 5.2.1 *A matrix-variate random variable $\Lambda \in \mathbb{S}_{++}^N$ is $\mathcal{MGIG}$ distributed [19, 32] and is denoted as $\Lambda \sim \mathcal{MGIG}_N(\Psi, \Phi, \nu)$ if the density of Λ is*

$$f(\Lambda) = \frac{|\ \Lambda\ |^{\nu-(N+1)/2}}{|\ \frac{\Psi}{2}\ |^{\nu}\ \ B_\nu(\frac{\Phi}{2}\frac{\Psi}{2})} \exp\{\mathrm{Tr}(-\frac{1}{2}\Psi\Lambda^{-1} - \frac{1}{2}\Phi\Lambda)\},$$

where $B_\nu(.)$ is the matrix Bessel function [87] defined as

$$B_\nu(\frac{\Phi}{2}\frac{\Psi}{2}) = |\frac{\Phi}{2}|^{-\nu}\int_{\mathbb{S}_{++}^N} |S|^{-\nu-\frac{N+1}{2}} \exp\{\mathrm{Tr}(-\frac{1}{2}\Psi S^{-1} - \frac{1}{2}\Phi S)\}dS. \tag{5.1}$$

The domain for parameters Φ *and* Ψ *for* $N \geq 2$ *is*

$$\begin{aligned} &\{\Psi \in \mathbb{S}_{+}^{N}, \Phi \in \mathbb{S}_{++}^{N}\} && \text{if } \nu \geq \frac{1}{2}N, \\ &\{\Psi \in \mathbb{S}_{++}^{N}, \Phi \in \mathbb{S}_{++}^{N}\} && \text{if } -\frac{1}{2}(N-1) \leq \nu < \frac{1}{2}N, \\ &\{\Psi \in \mathbb{S}_{++}^{N}, \Phi \in \mathbb{S}_{+}^{N}\} && \text{if } \nu < -\frac{1}{2}(N-1)\ . \end{aligned}$$

Next, we discuss special cases of $\mathcal{MGIG}$ distribution. When $N = 1$, the $\mathcal{MGIG}$ is the generalized inverse Gaussian distribution $\mathcal{GIG}$ [99] which is often used as the prior in several domains [23, 66]. If $\Psi = 0$, the $\mathcal{MGIG}$ distribution reduces to the Wishart, and if $\Phi = 0$, it becomes the Inverse Wishart distribution.

Proposition 13 *[229, Proposition 2] If matrix* $\Lambda \sim \mathcal{MGIG}_N(\Psi, \Phi, \nu)$*, then* $\Lambda^{-1} \sim \mathcal{MGIG}_N(\Phi, \Psi, -\nu)$.

Proof: The proof follows from the Bessel function property $B_\delta(WZ) = |WZ|^{-\delta} B_{-\delta}(ZW)$ [229]. ∎

Proposition 14 *If matrix* $\Lambda \sim \mathcal{MGIG}_N(\Psi, 0_N, \nu)$*, and* $-\nu > \frac{N-1}{2}$*, then* $\Lambda \sim IW_N(\Psi, -2\nu)$.

Proof: First note that If $-\nu > \frac{N-1}{2}$, then we have $B_\nu(0_N) = \Gamma_N(-\nu)$ [33]. Then, the proof simply follows from Definition 5.2.1. From Definition 5.2.1, the density of Λ is

$$f(\Lambda) = \frac{|\ \Lambda\ |^{\nu-(N+1)/2}}{|\ \frac{\Psi}{2}\ |^{\nu}\ \ B_\nu(0_N)} \exp\{\mathrm{Tr}(-\frac{1}{2}\Psi\Lambda^{-1})\} \tag{5.2}$$

$$= \frac{|\ \Lambda\ |^{\nu-(N+1)/2}}{|\ \frac{\Psi}{2}\ |^{\nu}\ \ \Gamma_N(-\nu)} \exp\{\mathrm{Tr}(-\frac{1}{2}\Psi\Lambda^{-1})\}, \tag{5.3}$$

which is the density function of $\Lambda \sim IW_N(\Psi, -2\nu)$. This completes the proof. ∎

Proposition 15 *If matrix* $\Lambda \sim \mathcal{MGIG}_N(0_N, \Phi, \nu)$*, and* $\nu > \frac{N-1}{2}$*, then* $\Lambda \sim W_N(\Phi^{-1}, 2\nu)$.

Proof: From Proposition 13, we have $\Lambda^{-1} \sim \mathcal{MGIG}_N(\Phi, 0_N, -\nu)$. Also, from Proposition 14, we have $\Lambda^{-1} \sim IW_N(\Phi, 2\nu)$. If matrix $\Lambda^{-1} \sim IW_N(\Phi, 2\nu)$ then $\Lambda \sim$

$W_N(\Phi^{-1}, 2\nu)$. This completes the proof. ∎

Sampling Mean of $\mathcal{MGIG}$: The sufficient statistics of $\mathcal{MGIG}$ are $\log|\Lambda|$, Λ, and Λ^{-1}. It is, however, difficult to analytically calculate the expectations $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda]$ and $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda^{-1}]$. Importance sampling can be applied to approximate those quantities. Note that based on the result of Proposition 13, the importance sampling procedure for estimating mean of $\mathcal{MGIG}$, i.e., $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda]$, can also be applied to infer the reciprocal mean i.e. $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda^{-1}]$.

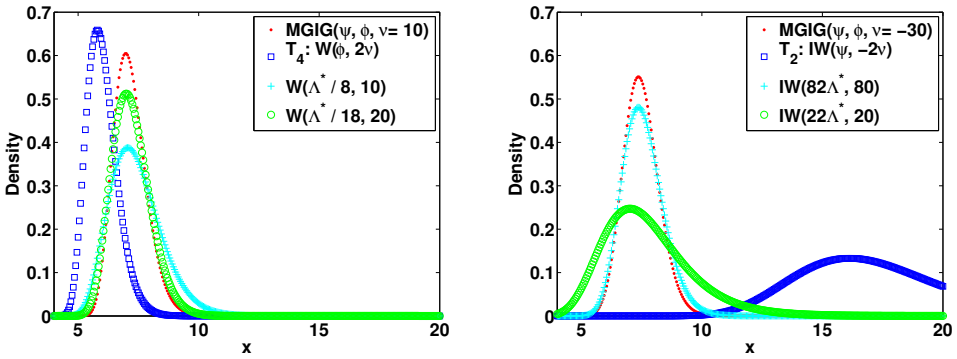
An importance sampling procedure proposed in [229, 233], where the $\mathcal{MGIG}$ is viewed as a product of Inverse Wishart and Wishart distributions and one of the multiplicands is used as the natural choice of the proposal distribution. In particular, in [229, 233], the $\mathcal{MGIG}$ is viewed as

$$\mathcal{MGIG}_N(\Lambda|\Psi,\Psi,\nu) \propto \underbrace{e^{\mathrm{Tr}(-\frac{1}{2}\Phi\Lambda)}}_{T_1}\underbrace{\mathcal{IW}_N(\Lambda\,|\,\Psi,-2\nu_u)}_{T_2} \propto \underbrace{e^{\mathrm{Tr}(-\frac{1}{2}\Psi\Lambda^{-1})}}_{T_3}\underbrace{\mathcal{W}_N(\Lambda\,|\,\Phi,2\nu_u)}_{T_4}.$$

In [229, 233], authors advocate using T_2 (or T_4) as the proposal distribution which simplify the weight calculation to the evaluation of T_1 (or T_3). However, it is not studied how close T_2 (or T_4) are to the $\mathcal{MGIG}$ distribution in shape. For example, consider the $1-$dimensional $\mathcal{MGIG}$ distribution

$$\mathcal{MGIG}_1(\Lambda\,|\,35,10,10) \propto \underbrace{e^{\mathrm{Tr}(-\frac{35}{2}\Lambda^{-1})}}_{T_3}\underbrace{\mathcal{W}_1(\Lambda\,|\,10,20)}_{T_4}. \tag{5.4}$$

In [229, 233], $T_4 : \mathcal{W}_1(\Lambda\,|\,10,20)$ is considered as the proposal distribution, but the mode of T_4 is far away from the mode of $\mathcal{MGIG}_1(\Lambda\,|\,35,10,10)$ (Figure 5.2(a)). As a result, samples drawn from T_4 will be on the tail of the $\mathcal{MGIG}_1(\Lambda\,|\,10,20)$ distribution, and will end up getting low weights from the $\mathcal{MGIG}_1(\Lambda\,|\,10,20)$ distribution. Such a sampling procedure will be wasteful, i.e., drawing samples from the tails of the target $\mathcal{MGIG}_1$ distribution, leading to a very low ESS. Similar behavior is observed with several different choices of parameters for the $\mathcal{MGIG}$, here we only show three of them in Figures 5.2 and 5.3 due to the lack of space.



(a) (b)

Figure 5.2: (a,b) Comparison of different proposal distribution (a) Wishart ($\mathcal{W}$) and (b) Inverse Wishart ($\mathcal{IW}$) for sampling mean of $\mathcal{MGIG}_1(\Psi, \Phi, \nu)$ where Λ^* is the mode of $MGIG$. The blue curves are the proposal distribution defined in [229, 233] which can not recover the mode of the $\mathcal{MGIG}$ distribution.

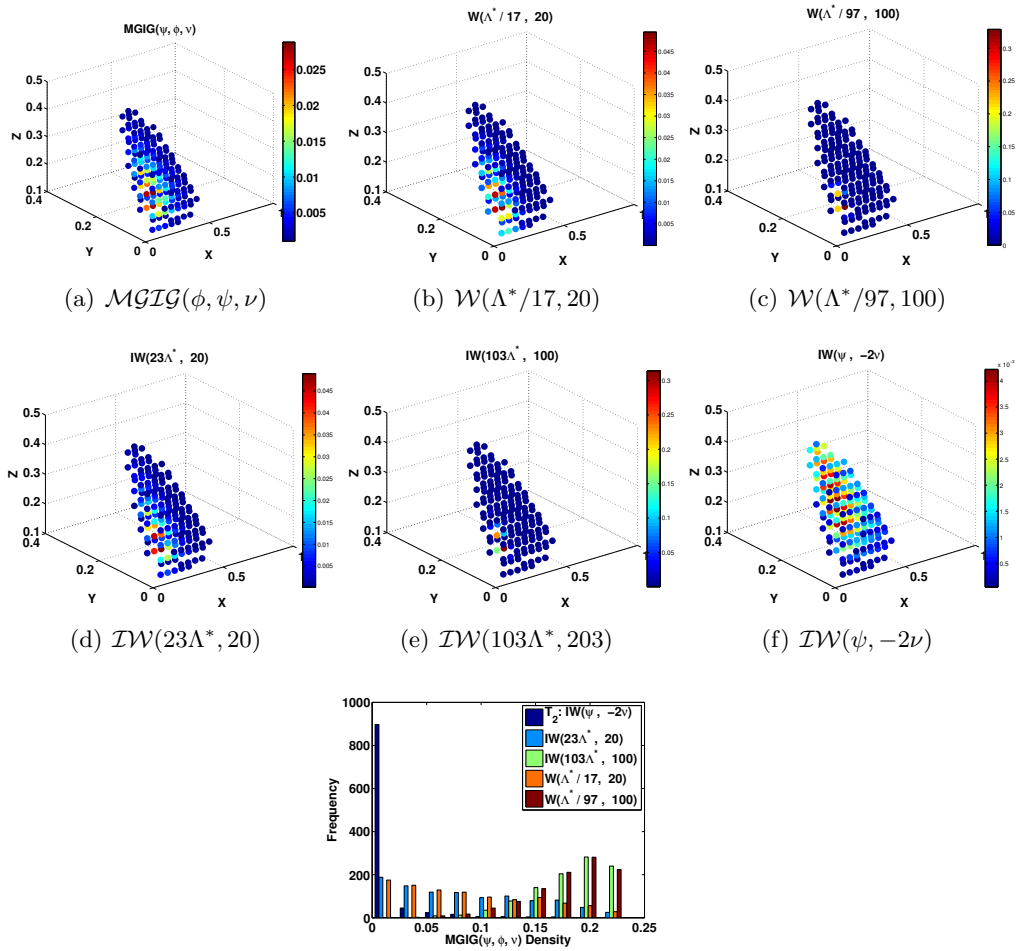
5.2.3 Algebraic Riccati Equation

An algebraic Riccati equation (ARE) is

$$A^T X + XA + XRX + Q = 0, \tag{5.5}$$

where $A \in \mathbb{R}^{N\times N}, Q \in \mathbb{S}_+^N$, and $R \in \mathbb{S}_+^N$. We associate a $2N \times 2N$ matrix called the Hamiltonian matrix H with the ARE (5.5) as $H = \begin{bmatrix} A & R \\ -Q & -A^T \end{bmatrix}$. The ARE (5.5) has a unique positive definite solution if and only if the associated Hamiltonian matrix H has no imaginary eigenvalues (Section 5.6.3 of [28]).

There have been offered various numerical methods to solve the ARE which can be reviewed in [9]. The key of numerical technique to solve ARE (5.5) is to convert the problem to a stable invariant subspace problem of the Hamiltonian matrix i.e., finding the invariant subspace corresponding to the eigenvalues of H with negative real parts. In particular, consider $V = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ to be a $H-$invariant subspace, i.e., $HV = V\Lambda$. Assume



(a) $\mathcal{MGIG}(\phi, \psi, \nu)$ (b) $\mathcal{W}(\Lambda^*/17, 20)$ (c) $\mathcal{W}(\Lambda^*/97, 100)$

(d) $\mathcal{IW}(23\Lambda^*, 20)$ (e) $\mathcal{IW}(103\Lambda^*, 203)$ (f) $\mathcal{IW}(\psi, -2\nu)$

(g)

Figure 5.3: Illustration of 2-dimensional (a) $\mathcal{MGIG}$ distribution (b-f) and different proposal distributions where (b-e) are the proposal described in this chapter where Λ^* is the mode of $MGIG$ and (f) is the proposal defined in [229, 233]. the proposal distribution defined in [229, 233] (f) can not recover the mode of the $\mathcal{MGIG}$ distribution (a). (g) Density of $\mathcal{MGIG}_2(\Psi, \Phi, \nu)$ for 1000 samples generated by each proposal distribution is calculated. More than 90% of samples generated by the previous proposal distribution in [229, 233] ($\mathcal{IW}(\psi, -2\nu)$) have zero $\mathcal{MGIG}$ density leading to $ESS = 40$. Whereas, the new proposal distribution $IW(23\Lambda^*, 20)$ has the $ESS = 550$ which has a very similar shape to the target $\mathcal{MGIG}$ distribution.

X_1 is invertible, we then post multiply by X_1^{-1} to obtain

$$HVX_1^{-1} = V\Lambda X_1^{-1} \Rightarrow \begin{bmatrix} A & R \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} X_1 \Lambda X_1^{-1}, \tag{5.6}$$

where $X = X_2 X_1^{-1}$. Multiplying both side by $\begin{bmatrix} -X & I \end{bmatrix}$, we have

$$\begin{bmatrix} -X & I \end{bmatrix} \begin{bmatrix} A & R \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} -X & I \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} X_1 \Lambda X_1^{-1} = 0. \tag{5.7}$$

Simplifying the left hand side we get the ARE (5.5) which implies that $X = X_2 X_1^{-1}$ is the solution of (5.5). The usual ARE solvers such as the Schur vector method [114], SR methods [31], the matrix sign function [11, 34] require in general $O(n^3)$ flops [123]. For special cases, faster algorithms such as [123] can be applied which solves such an ARE with 20k dimensions in seconds. In this chapter, we use Matlab ARE solver (*care*) to find the solution of ARE.

5.3 MGIG Properties and Sampling

Some properties of the $\mathcal{MGIG}$ distribution and its connection with Wishart distribution has been studied in [32, 180, 181]. However, to best of our knowledge, it is not yet known if the distribution is unimodal and how to obtain the mode of $\mathcal{MGIG}$. In the following Lemma we show that the $\mathcal{MGIG}$ distribution is unimodal.

Lemma 16 *Consider the $\mathcal{MGIG}$ distribution $\mathcal{MGIG}_N(\Lambda|\Psi, \Phi, \nu)$. The mode of $\mathcal{MGIG}$ distribution is the solution of the following Algebraic Riccati Equation (ARE)*

$$-2\alpha\Lambda + \Lambda\Phi\Lambda - \Psi = 0, \tag{5.8}$$

where $\alpha = (\nu - \frac{N+1}{2})$. ARE in (5.8) has a unique positive definite solution, thus the $\mathcal{MGIG}$ distribution is a unimodal distribution.

Importance Sampling for $\mathcal{MGIG}$: Since $\mathcal{MGIG}$ is a unimodal distribution, we propose an efficient importance sampling procedure for $\mathcal{MGIG}$ by mode matching. We

select a proposal distribution over space of positive definite matrices by matching the proposal's mode to the mode of $\mathcal{MGIG}$ (mode matching) which aligns the proposal and $\mathcal{MGIG}$ shapes. Mode matching is a good choice of the proposal as the proposal $q(x)$ is large in a region where the target distribution $\mathcal{MGIG}$ is large leading to a good estimate of the expectations $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda]$ or $\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[\Lambda^{-1}]$. An example of such proposal distribution is Inverse Wishart or Wishart distribution.

Let Λ^* be the mode of $\mathcal{MGIG}_N(\Lambda|\Psi,\Phi,\nu)$ which can be found by solving the ARE (5.8). The mode of Inverse Wishart $\mathcal{W}_N(\Lambda|\Sigma,\rho)$ distribution is $\Sigma^* = (\rho - N - 1)\Sigma$. To match the mode of $\mathcal{W}_N(\Lambda|\Sigma,\rho)$ with that of $\mathcal{MGIG}_N(\Lambda|\Psi,\Phi,\nu)$, we choose the scale parameter Σ of the Wishart distribution by setting $\Sigma^* = \Lambda^*$. In particular,

$$\Sigma^* = \Lambda^* = (\rho - N - 1)\Sigma \quad \Rightarrow \quad \Sigma = \frac{\Lambda^*}{\rho - N - 1}. \tag{5.9}$$

Thus, we suggest using $\mathcal{W}_N(\frac{\Lambda^*}{\rho-N-1},\rho)$ as the proposal distribution. At each iteration, we draw a sample $\Lambda_i \sim \mathcal{W}_N(\frac{\Lambda^*}{\rho-N-1},\rho)$, and calculate the importance weight as

$$w(\Lambda_i) = \frac{\mathcal{MGIG}_N(\Lambda_i|\Psi,\Phi,\nu)}{\mathcal{W}_N(\Lambda_i|\Sigma,\rho)} = \mid \Lambda_i \mid^{\nu-\frac{\rho}{2}} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\Psi\Lambda_i^{-1} + [\Phi - \Sigma^{-1}]\Lambda_i\right)\right\}.$$

As a result, we can approximate the sample mean as

$$\mathbb{E}_{\Lambda\sim\mathcal{MGIG}}[f(\Lambda)] = \frac{\sum_{i=1}^{S} w(\Lambda_i) f(\Lambda_i)}{\sum_{j=1}^{S} w(\Lambda_j)}. \tag{5.10}$$

Note that the weight calculation requires to calculate the inverse and determinant of sampled matrix Λ_i. However, the random samples generator from $\mathcal{W}$ [189] returns the upper triangular matrix R where $\Lambda = R^T R$. Hence the inverse and determinant of Λ can be calculated efficiently from the inverse and diagonal of the triangular matrix R, respectively. Therefore, the cost of weight calculation is reduced to the cost of solving a linear system and upper triangular matrix production at each iteration.

A similar argument holds when the proposal distribution is an Inverse Wishart distribution. In particular, the mode of Inverse Wishart $\mathcal{IW}_N(\Sigma,\rho)$ distribution is $\frac{\Sigma}{\rho+N+1}$. Thus $\mathcal{IW}_N(\rho+N+1)\Lambda^*,\rho)$ is another suitable choice of the proposal distribution.

Figure 5.2 illustrates that the proposed importance sampling outperforms the one in

[229, 233] for three examples of $\mathcal{MGIG}$. In particular, more than 90% of samples drawn from the proposal distribution T_2 in [229, 233] have zero weights leading to $ESS = 40$ (Figure 5.2 (c)). Whereas, our proposal distribution achieved $ESS = 550$ leading to a better approximation of the mean of $\mathcal{MGIG}$. Similar behavior is observed with several different choices of parameters for the $\mathcal{MGIG}$.

5.4 Connection of MGIG and Bayesian PCA

In this section, we illustrate that the mapping matrix V in Bayesian PCA can be marginalized or 'collapsed' yielding a Matrix Generalized Inverse Gaussian ($\mathcal{MGIG}$) [19, 32] posterior distribution over the latent matrix U denoting as the marginalized posterior distribution. Then, we explain the derivation of the marginalized posterior for data with missing values, followed by a collapsed Monte Carlo Inference for PMF.

5.4.1 Closed form Posterior Distribution in Bayesian PCA

The key challenge in models such as Bayesian PCA or Bayesian PMF is that joint marginalization over both latent factors U, V is intractable. Probabilistic PCA gets around the problem by considering one of the variables, say V, to be a constant. In this section, we show that one can marginalize or 'collapse' one of the latent factors, say V, and obtain the marginalized posterior $P(U|X)$ over the other variable denoted. In fact, we obtain the posterior with respect to the covariance structure $\Lambda_u = \beta_u \mathbb{I} + UU^T$, for a suitable constant β_u, which is sufficient to do Bayesian inference on new test points x_{test}.

We start with an outline of the analysis. Note that

$$p(U|X) \propto p(U)P(X|U) = p(U) \int_V P(X|U,V)p(V)dV \ , \tag{5.11}$$

and, based on the posterior over U, one can obtain the probability on a new point as

$$p(x_{\text{test}}|X) = \int_U p(x_{\text{test}}|U)p(U|X)dU \ . \tag{5.12}$$

Next, we show that the posterior over U as in (5.11), rather the distribution over $\Lambda_u = \beta_u \mathbb{I} + UU^T$, can be derived analytically in *closed form*. The distribution is the

Matrix Generalized Inverse Gaussian ($\mathcal{MGIG}$) distribution.

Now, similar to (2.5), marginalizing V gives

$$p\left(X \mid U\right) = \int_V p(X|U,V)p(V)dV = \prod_{m=1}^{M} \mathcal{N}\left(\mathbf{x}_{:m} \mid 0, \sigma_v^2 \Lambda_u\right), \tag{5.13}$$

where $\Lambda_u = \beta_v \mathbb{I} + UU^T$ and $\beta_v = \frac{\sigma^2}{\sigma_v^2}$. Then, the marginalized posterior of U is calculated as

$$\begin{aligned} p(U \mid X) \propto p(X|U)\ p(U) \quad &\propto \quad \mid \Lambda_u \mid^{-M/2} \exp\left\{ \frac{-\operatorname{Tr}\left(\Lambda_u^{-1} \sum_{m=1}^{M} \mathbf{x}_{:m}\mathbf{x}_{:m}^T\right)}{2\sigma_v^2} \right\} \\ &\qquad \times \exp\left\{ \frac{-\operatorname{Tr}(\Lambda_u)}{2\sigma_u^2} \right\} \times \exp\left\{ \frac{\operatorname{Tr}(\beta_u \mathbb{I})}{2\sigma_u^2} \right\} \qquad (5.14) \\ &= \quad \mid \Lambda_u \mid^{-M/2} \exp\left\{ \operatorname{Tr}(-\frac{1}{2}\Lambda_u^{-1}\Psi_u - \frac{1}{2}\Lambda_u \Phi_u) \right\} \\ &\sim \quad \mathcal{MGIG}(\Lambda_u \mid \Psi_u, \Phi_u, \nu_u)\ , \qquad (5.15) \end{aligned}$$

where $\Psi_u = \frac{1}{\sigma_v^2} XX^T$, $\Phi_u = \frac{1}{\sigma_u^2}\mathbb{I}$, and $\nu_u = \frac{N-M+1}{2}$.

Therefore, by marginalizing or collapsing V, the posterior over $\Lambda_u = \beta_v \mathbb{I} + UU^T$ corresponding to the latent matrix U can be characterized exactly with a $\mathcal{MGIG}$ distribution with parameters depending only on X. Note that this is in sharp contrast with (2.6) for PPCA, where the posterior covariance of $\mathbf{u}_n$ is $\sigma^{-2}\Gamma$ which in turn depends on the point estimate for $\hat{V}$.

5.4.2 Posterior Distribution with Missing Data

In this section, we consider the matrix completion setting, when the observed matrix X has missing values. In presence of missing values, the likelihood of the observed sub-vector in any column of X is given as

$$p\left(\mathbf{x}_{n_m,m} \mid U, V\right) = \mathcal{N}\left(\mathbf{x}_{n_m,m} \mid \tilde{U}_m \mathbf{v}_m^T, \sigma^2 \mathbb{I}\right). \tag{5.16}$$

where n_m is a vector of size $\tilde{N}_m$ containing indices of non-missing entries in column m of X, and $\tilde{U}_m$ is a sub-matrix of U with size of $\tilde{N}_m \times D$ where each row correspond to a non-missing entry in the m^{th} column of X. The marginalized likelihood (5.13) can be

written as

$$p\left(X \mid U\right) = \prod_{m=1}^{M} \mathcal{N}\left(\mathbf{x}_{n_m,m} \mid 0,\ \sigma_v^2 \Lambda_{un}\right), \tag{5.17}$$

where $\Lambda_{un} = \beta_v \mathbb{I} + \tilde{U}_n \tilde{U}_n^T$ and $\beta_v = \frac{\sigma^2}{\sigma_v^2}$. The marginalized posterior is given by

$$p(U \mid X) \propto \times \mid \Lambda_{un} \mid^{-M/2} \exp\left\{-\frac{1}{2}\mathbf{x}_{n_m,m}^T\ \Lambda_{un}^{-1}\ \mathbf{x}_{n_m,m}\right\} \exp\left\{-\frac{1}{2\sigma_u^2}\ \mathrm{Tr}(UU^T)\right\}\ .$$

As shown in (5.18), in presence of missing values, the posterior cannot be factorized as in (5.14) because each column $\mathbf{x}_{:m}$ contributes to different blocks Λ_{un} of Λ.

We propose to address the missing value issue by gap-filling. In particular, if one can obtain a good estimate of the covariance structure in X, so that $\Psi_u = \frac{1}{\sigma_v^2} XX^T$ in (5.15) can be approximated well, one can use the $\mathcal{MGIG}$ posterior to do approximate inference. We consider two simple approaches to approximate the covariance structure of X: (i) by zero-padding the missing value matrix X (assuming $E[X] = 0$ or centering the data in practice), and estimating the covariance structure based on the zero-padded matrix, and (ii) by using a suitable matrix completion method, such as PMF, to get point estimates of the missing entries in X, and estimating the covariance structure based on the completed matrix. We experiment with both approaches in Section 6.3.4, and the zero-padded version seems to work quite well.

5.4.3 Collapsed Monte Carlo Inference for PMF

Given that $\Lambda_u \sim \mathcal{MGIG}_N$, we predict the missing values as follows. Let $\mathbf{x} = [\mathbf{x}^o, \mathbf{x}^*] \sim \mathcal{N}(0, \Lambda)$, where $\mathbf{x}^o \in \mathbb{R}^p$ is the observed partition of $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{x}^* \in \mathbb{R}^{N-p}$ is missing. Accordingly, partition Λ as

$$\Lambda_u = \begin{array}{c} \begin{array}{cc} p & N-p \end{array} \\ \begin{pmatrix} \Lambda_{oo} & \Lambda_{o*} \\ \Lambda_{*o} & \Lambda_{**} \end{pmatrix} \end{array} \begin{array}{c} p \\ N-p \end{array}\ . \tag{5.18}$$

Algorithm 3 CMC Inference for PMF

1: Construct zero-padded matrix Z from $X \in \mathbb{R}^{N\times M}$.
2: Let $\Psi_u = \frac{ZZ^T}{\sigma_v^2}$, $\Phi_u = \frac{\mathbb{I}}{\sigma_u^2}$, and $\nu_u = \frac{N-M+1}{2}$.
3: Solve (5.8) to find mode Λ^* of $\mathcal{MGIG}(\Psi_u, \Phi_u, \nu_u)$.
4: Let $L^T L = \Lambda^*$ be the Cholesky factorization of Λ^*. Let $\tilde{L} = \frac{L}{\sqrt{\rho - M - 1}}$.
5: **for** $t = 1 \cdots T$ **do**
6: Let $\Lambda^{(t)} \sim \mathcal{W}_N(\frac{\Lambda^*}{\rho-M-1}, \rho, \tilde{L})$ ▷ Algorithm 1
7: Let $w^t = \frac{\mathcal{MGIG}_N(\Lambda^{(t)}|\Psi_u,\Phi_u,\nu_u)}{\mathcal{W}_N(\Lambda^{(t)}|\frac{\Lambda^*}{\rho-M-1},\rho,\tilde{L})}$.
8: Let $\mu^t = \Lambda_{*o}^{(t)} {\Lambda_{oo}^{(t)}}^{-1} \mathbf{x}^o$. Let $\Sigma^t = \Lambda_{**}^{(t)} - \Lambda_{*o}^{(t)} {\Lambda_{oo}^{(t)}}^{-1} \Lambda_{o*}^{(t)}$.
9: Let $\bar{\mu} = \bar{\mu} + w^t \mu^t$. Let $\bar{\Sigma} = \bar{\Sigma} + w^t \Sigma^t$.
10: Report the distribution of $\mathbf{x}^* \sim \mathcal{N}(\tilde{\mu}^*, \tilde{\Sigma}^*)$ where $\tilde{\mu}^* = \frac{\bar{\mu}}{\sum_{t=1}^T w^t}$ and $\tilde{\Sigma}^* = \frac{\bar{\Sigma}}{\sum_{t=1}^T w^t}$.

Then, the conditional probability of $\mathbf{x}^*$ given $\mathbf{x}^o$ and Λ is

$$p(\mathbf{x}^* \,|\, \mathbf{x}^o, \Lambda) \sim \mathcal{N}(\mu^*, \Sigma^*), \qquad \mu^* = \Lambda_{*o}\Lambda_{oo}^{-1}\mathbf{x}^o, \quad \Sigma^* = \Lambda_{**} - \Lambda_{*o}\Lambda_{oo}^{-1}\Lambda_{*o}.$$

where $\mathbf{y} = \Lambda_{*o}\Lambda_{oo}^{-1}$ is the solution of the linear system $\Lambda_{oo}\mathbf{y} = \Lambda_{*o}^T$ and can be calculated efficiently. Since sampling from $\mathcal{MGIG}$ is difficult, we propose to use importance sampling to infer the missing values as

$$p(x_n^*|x_n^o) = \mathbb{E}_{\Lambda\sim\mathcal{MGIG}}\left[p(x_n^*|x_n^o,\Lambda)\right] = \mathbb{E}_{\Lambda\sim q}\left[\frac{p(x_n^*|x_n^o,\Lambda)\mathcal{MGIG}_N(\Lambda|\Psi_u,\Phi_u,\nu_u)}{q(\Lambda)}\right],$$

where q is the proposal distribution as discussed above and sampling $\Lambda^{(t)}$ from q yields to the estimate of

$$\tilde{\mu}^* = \frac{\sum_{t=1}^T \Lambda_{*o}^{(t)} {\Lambda_{oo}^{(t)}}^{-1} \mathbf{x}^o w(\Lambda^{(t)})}{\sum_{t=1}^t w(\Lambda^{(t)})}, \qquad \tilde{\Sigma}^* = \frac{\sum_{t=1}^T [\Lambda_{**}^{(t)} - \Lambda_{*o}^{(t)} {\Lambda_{oo}^{(t)}}^{-1} \Lambda_{*o}^{(t)}] w(\Lambda^{(t)})}{\sum_{t=1}^t w(\Lambda^{(t)})}. \quad (5.19)$$

Algorithm 3 illustrates the summary of the collapsed Monte Carlo (CMC) inference for predicting the missing values. A practical approximation to avoid the calculations in Lines 9-12 of Algorithm 3 at each iteration is to simply estimate the mean of the posterior $\bar{\Lambda} = \frac{\sum_{t=1}^T \Lambda^{(t)} w^t}{\sum_{t=1}^T w^t}$ with samples drawn from the proposal distribution (line 6), then do the inference based on $\bar{\Lambda}$. As it is shown in Section 6.3.4, if the degrees of freedom ν_u is small, the mode is close to the mean and the approximation using $\bar{\Lambda}$

works well.

5.5 Experimental Results

We compared the performance of MCMC and CMC on both log loss and running times.

5.5.1 Datasets

We evaluated the models on 4 datasets:

SNP: Single nucleotide polymorphism (SNP) is important for identifying gene-disease associations where the data usually has 5 to 20% of genotypes missing [29]. We used phased SNP dataset for chromosome 13 of the CEU population[1]. We randomly dropped 20% of the entries.

Gene Expression: DNA microarrays provides measurement of thousands of genes under a certain experimental condition where suspicious values are usually regarded as missing values. Here we used gene expression dataset for Breast Cancer (BRCA) generated by the TCGA Research Network[2] with $M = 17,814$ genes and $N = 591$ samples. We randomly hide 20% of entries.

MovieLens: we used MovieLens[3] dataset with 1M rating represented as a fat matrix $X \in \mathbb{R}^{N \times M}$ where $M = 3900$ movies and $N = 6040$ users. Movies with less than 10 ratings are removed yielding to $M = 3233$ movies. We randomly selected 100 movies among 3900 movies in the dataset. The used subset of MovieLens contains 83,000 ratings meaning 86% of the ratings are missing.

Synthetic: first the latent matrices U and V are generated by randomly choosing each $\{\mathbf{u}_n\}_{n=1}^N$ and $\{\mathbf{v}_m\}_{m=1}^M$ from $\mathcal{N}(0, \sigma_u^2 \mathbb{I})$ and $\mathcal{N}(0, \sigma_v^2 \mathbb{I})$, respectively. Then, matrix X is built by sampling each x_{nm} from $\mathcal{N}(\langle \mathbf{u}_n, \mathbf{v}_m \rangle, \sigma^2)$. The parameters are set to $N = 100$, $M = 6000$, $\sigma_u^2 = \sigma_v^2 = 0.05$, and $\sigma^2 = 0.01$. We dropped random entries using Bernoulli distributions with $\delta = 0.1, 0.2$.

[1] http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/

[2] http://cancergenome.nih.gov/

[3] www.movielens.umn.edu

5.5.2 Methodology

We compared CMC with MCMC inference for PMF. Gibbs sampling with diagonal covariance prior over the latent matrices is used for MCMC. For the model evaluation, average of log loss (LL) is reported over 5-fold cross-validation. LL measures how well a probabilistic model q predicts the test sample defined as $LL = -\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{ij}\log q(x_{ij})$ where $q(x_{ij})$ is the inferred probability and T is the total number of observed values. A better model q assign higher probability $q(x_{ij})$ to observed test data, and have a smaller value of LL.

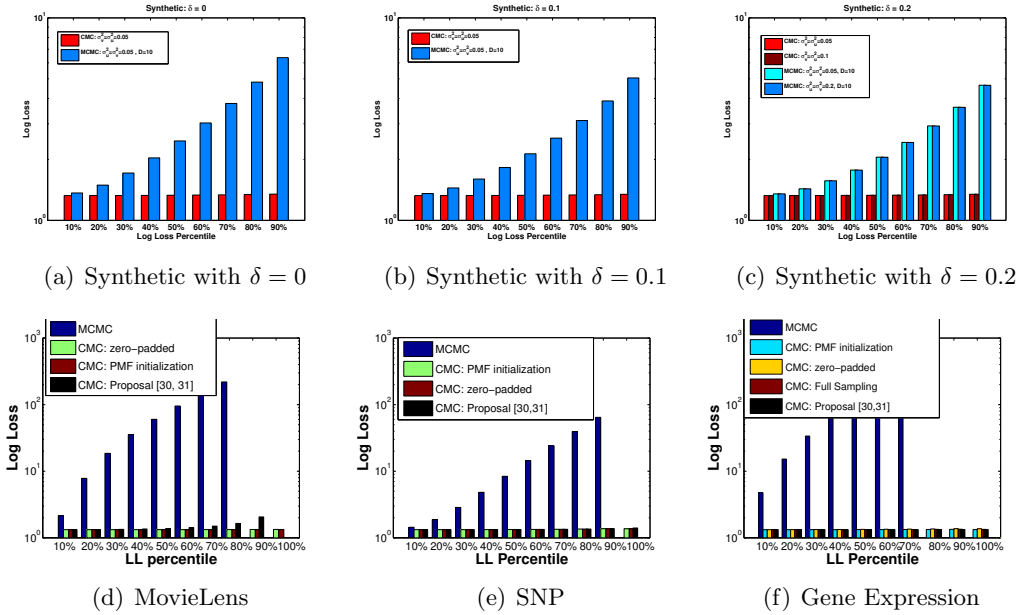
LL Percentile: For any posterior model $q(x)$, a test data point x_{test} with low $q(x_{\text{test}})$ has large log loss, and high $q(x_{\text{test}})$ has low log loss. To comparatively evaluate the posteriors obtained from MCMC and CMC, we consider their log loss percentile plots. For any posterior, we sort all the test data points in ascending order of their log loss, and plot the mean log loss in 10 percentile batches. More specifically, the first batch corresponds to the top 10% of data points with the lowest log loss, the second batch corresponds to the top 20% of data points with the lowest log loss (including the first 10% percentile), and so on.

5.5.3 Results

We summarize the results from different aspects: First, we we comparatively evaluate CMC and MCMC based on log loss, then the effective number of samples used in CMC and MCMC is discussed. We show the affect of the different initialization methods, and compare the full sampler and vs mean sampler. Finally, the inferred posterior distributions from CMC and MCMC are compared, and time comparison of algorithms is provided.

Log Loss

CMC has a small log loss across all percentile batches, whereas log loss of MCMC increases exponentially (linear increase in the log scale) for percentile batches with higher log loss i.e., smaller predicting probability, (Figure 5.4). Thus, MCMC assigned extremely low probability to several test points as compared to CMC. Figure 5.5(a) illustrates that log loss of MCMC continues to decrease with growing sample size up to



(a) Synthetic with $\delta = 0$ (b) Synthetic with $\delta = 0.1$ (c) Synthetic with $\delta = 0.2$

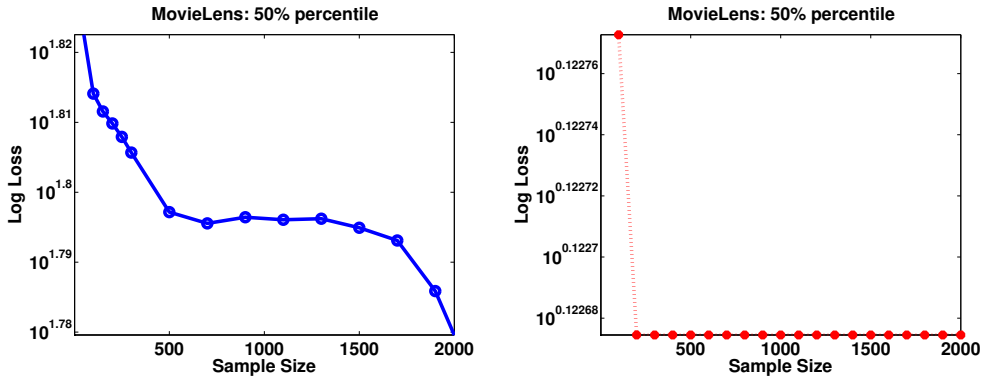
(d) MovieLens (e) SNP (f) Gene Expression

Figure 5.4: Log loss (LL) of CMC and MCMC for different log loss percentile on different datasets presented in the log scale (δ denotes the missing proportion). CMC consistently achieves lower LL compared to MCMC. LL of MCMC increases exponentially (linearly in log scale) by adding data points with higher log loss. Proposal in [30,31] achieved infinity LL for MovieLens. Empty bar represents infinity LL (e.g. 90% and 100% percentile in (d)

2000 samples, implying that MCMC has not yet converged to the equilibrium distribution. Note that log loss of CMC with 200 samples (Figure 5.5(b)) is 10 times less than log loss of MCMC with 2000 samples. We also compared the results with the previous proposal [229, 233], and observed that for MovieLens the results are worse than our proposed result as they achieved Inf LL on the last batch.

Effective Number of Samples

For the synthetic, SNP, and gene expression datasets, we generated 10,000 samples using MCMC. The burn-in period is set to 500 with a lag of 10 yielding to 1000 effective samples. For the MovieLens, we generated 5,000 samples using MCMC with the burn-in period of 1000 and a lag of 2 yielding to 2000 effective samples. We initialized the latent matrices U and V with the factors estimated by PMF, to help the convergence



(a) LL of MCMC (b) LL of CMC

Figure 5.5: LL of CMC and MCMC for different sample size of MovieLens data in the log scale. LL of both CMC and MCMC is decreasing by adding more samples. LL of MCMC is in magnitude 10 times more than CMC.

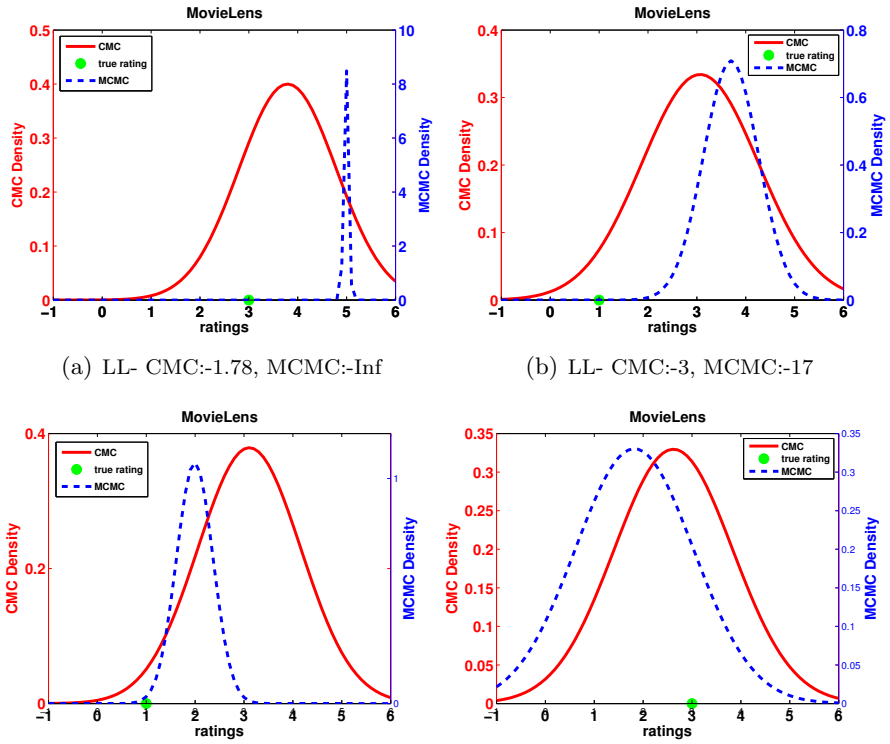
of MCMC. Sample size in CMC procedure is set to 1,000 for all datasets. Note that MCMC alternately sample both latent matrices U and V from a Markov chain and the quality of the posterior improves with increasing number of samples. For the proposed CMC procedure, the bigger matrix V is marginalized and only samples from the smaller U matrix is drawn directly from the true posterior distribution. Hence, CMC has considerably improved sample utilization.

Initialization

As discussed in Section 5.4.2, in order to use the $\mathcal{MGIG}$ posterior for inference, the covariance structure of matrix X should be estimated. Here we evaluate two approaches to approximate the covariance structure of X: (i) by zero-padding the missing value matrix X, and (ii) by computing the point estimates of the missing entries in X with PMF. CMC with zero-padded initialization has a similar log loss behavior as point estimate initialization with PMF (Figures 5.4 (d-f)).

Full Sampler vs Mean Sampler

Figure 5.4(f) shows the result of the full sampler (Algorithm 3), and the mean sampler (approximating the inference by estimating $\bar{\Lambda} = \mathbb{E}_{\Lambda \sim \mathcal{MGIG}}[\Lambda]$ as discussed in Section



(a) LL- CMC:-1.78, MCMC:-Inf

(b) LL- CMC:-3, MCMC:-17

(c) LL- CMC:-4.2, MCMC:-6.4

(d) LL- CMC:-1.4, MCMC:–2.04

Figure 5.6: Density of CMC and MCMC for several data input on MovieLens data. CMC provide distributions with lower LL compared to MCMC e.g. in (a) LL of MCMC is -Inf whereas LL of CMC is -1.78.

6.2.4) on gene expression data. Since the log losses are similar with both samplers, and the behavior is typical, we presented log loss results on the other datasets only based on the mean sampler, which is around 100 times faster.

Comparison of Inferred Posterior Distributions

To emphasize the importance of choosing the right measure for comparison, e.g., log loss vs RMSE, we illustrate the inferred posterior distributions over several missing entries/ratings in MovieLens obtained from MCMC and CMC in Figure 5.6. Note that the scales for CMC (red) and MCMC (blue) are different. Overall, the posterior from CMC tends to be more conservative (not highly peaked), and obtains lower log loss across a range of test points. Interestingly, as shown in Figure 5.6(a), MCMC can make

Table 5.1: Time Comparison of CMC and MCMC on different datasets. At each step of MCMC, rows of U and V can be sampled in parallel denoted by MCMC parallel. The running time is reported over 1000 steps for both methods where MCMC has 200 effective samples and CMC has 1000 effective samples. Note that the effective number of samples of MCMC is less than 1000 and more steps is required to obtain enough samples. The number of iterations for convergence of CMC is much less than 1000 (Figure 5.5).

Dataset	**Size**	**MCMC**		**CMC**	
		Serial	**Parallel**	**Serial**	**Parallel**
Synthetic	100× 6,000	728s	404s	6s	4s
SNP	120×104,868	12,862s	5,859s	75s	22s
Gene Expression	591× 17,814	3,478s	2,278s	140s	90s
MovieLens	3,233× 6,040	2,350s	2,100s	5,387s	2,058s

mistakes with high confidence, i.e., predicts 5 stars with a peaked posterior whereas the true rating is 3 stars. Such troublesome behavior is correctly assessed with log loss, but not by RMSE since it does not consider the confidence in the prediction. As shown in Figure 5.6(d), for some test points, both MCMC and CMC inferred similar posterior distributions with a bias difference where the mean of CMC is closer to the true value.

Time Comparison

We have compared running time in both serial and in parallel over 1000 steps yielding to 200 and 1000 samples for MCMC and CMC, respectively. We implement the algorithms in Matlab. The computation time is estimated on a PC with a 3.40 GHz Quad core CPU and 16.0G memory. The average run time results are reported in Table 5.1. For Synthetic, SNP, and gene expression datasets, MCMC converges very slowly. For MovieLens dataset, the running time of both are very close but note that MCMC requires more number of samples for convergence than CMC (Figure 5.5).

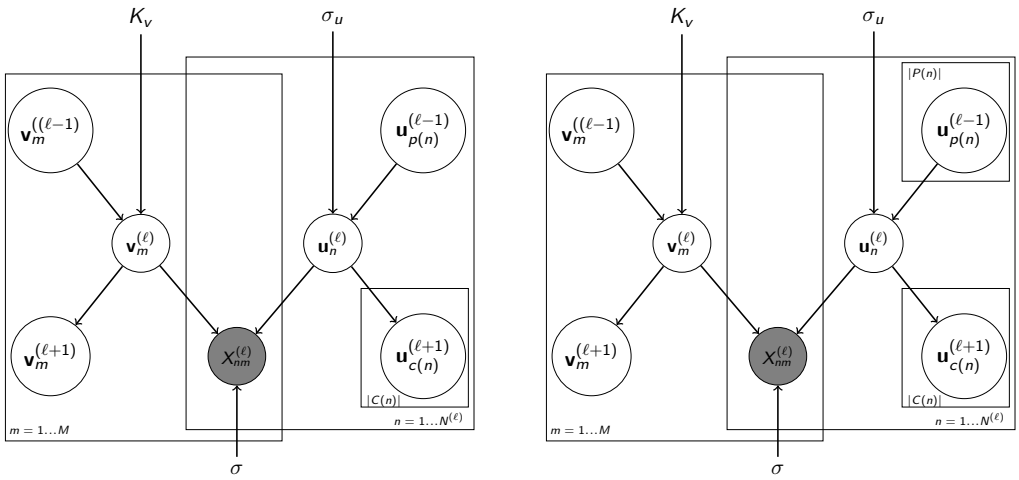
Chapter 6

Matrix Completion with Hierarchical Side Information

6.1 Introduction

A key limitation of most matrix factorization (MF) models is the inability to use the domain knowledge such as hierarchical side information. In fact, applying PMF model [174] which does not incorporate the plant taxonomic hierarchy leads to a performance worse than the simple algorithm MEAN which uses the domain knowledge [183] (Section 6.3.4). In a recent work, the hierarchical information is incorporated into MF in three different ways – hierarchical regularization, agglomerate fitting, and residual fitting [141]. In another work, hierarchical PMF (HPMF) was proposed for predicting missing values where inference in the model was done using alternating optimization leading to a point estimate of the missing values [183].

One of the main drawbacks of point estimates in the context of matrix completion is that they provide no uncertainty quantification. In most scientific disciplines, an uncertainty quantified prediction is essential for understanding the predictions, and planning subsequent steps, including additional data collection efforts to reduce uncertainties. In some applications like plant trait prediction, we are therefore interested in inferring a distribution for each prediction, which motivates the use of a Bayesian approach to the problem. BPMF is proposed in [175] as a Bayesian generalization of PMF by maintaining a distribution over all possible covariance matrices. Gibbs sampler is applied



(a) BHPMF (b) MI-BHPMF

Figure 6.1: (a) BHPMF and (b) MI-BHPMF schematic at level (ℓ). In spite of the size of the model, the Gibbs sampler is efficient since the Markov blanket is small and independent of the number of levels. MI-BHPMF supports multiple inheritance.

for inference in BPMF yielding to a distribution for each prediction.

In this chapter, we present Bayesian HPMF (BHPMF) model along with inference algorithms for uncertainty quantified matrix completion while incorporating a given hierarchy. The development is more general than that in [141, 183] since we provide uncertainty quantified predictions The model is also more general than that in [175] since we incorporate the given hierarchical side information as part of the model.

The rest of chapter is organized as follows. Section 6.2 presents the BHPMF model. We present the experimental results for plant trait prediction in Section 6.3. In Section 6.4, we introduce a hierarchical multiple inheritance model with the preliminary results on movie recommendation.

6.2 BHPMF

In this Section, we propose a full Bayesian model (BHPMF) and an inference procedure that incorporates the hierarchical side information and provides uncertainty quantified estimates of the missing trait values.

6.2.1 Model specification

We illustrate the BHPMF model in the context of plant trait prediction. However, it can be applied to any problem with the hierarchical side information.

Denote the data matrix at each level ℓ with $X^{(\ell)} \in \mathbb{R}^{N^{(\ell)} \times M}$ for ℓ running from the top level 1 (e.g. phylogenetic groups) to the bottom level L (e.g. individual plants). Each row $n^{(\ell)}$ and column $m^{(\ell)}$ of $X^{(\ell)}$ has a latent factor $\mathbf{u}_n^{(\ell)} \in \mathbb{R}^D$ and $\mathbf{v}_m^{(\ell)} \in \mathbb{R}^D$, respectively. Denoting latent factor matrices at level ℓ with $U^{(\ell)} \in \mathbb{R}^{N^{(\ell)} \times D}$ and $V^{(\ell)} \in \mathbb{R}^{M \times D}$ (Figure 6.1(a)).

The generative process of BHPMF at level ℓ is given as follows:

1. Generate $\mathbf{u}_n^{(\ell)} \sim N(\mathbf{u}_{p(n)}^{(\ell-1)}, \sigma_u^2 I)$, $[n]_1^{N^{(\ell)}}$, where $p(n)$ is the parent node of n in the upper level.
2. Generate $\mathbf{v}_{:d}^{(\ell)} \sim N(\mathbf{v}_{:d}^{(\ell-1)}, [K_v^{(\ell)}]^{-1})$, $[d]_1^D$.
3. Generate $x_{nm}^{(\ell)} \sim N(\langle \mathbf{u}_n^{(\ell)}, \mathbf{v}_m^{(\ell)} \rangle, \sigma^2)$ for each non-missing entry, where $\langle ., . \rangle$ is the inner product.

where $\mathbf{v}_{:d}^{(\ell)} \in \mathbb{R}^M$ is column d of $V^{(\ell)}$ and $K_v^{(\ell)}$ is the trait precision matrix (inverse of covariance matrix) at level ℓ.

We use a Gibbs sampling procedure [81] to draw samples of latent matrices from the joint posterior. In spite of the size of the model, the Gibbs sampler is efficient since the Markov blanket is small and independent of the number of levels (Figure 6.1(a)). We consider two different types of trait covariance structure prior for trait factors: diagonal covariance and full covariance matrix.

6.2.2 Sampling U

Let $C(n) = \{c_i(n)\}$ be a set of child nodes of n with $c_i(n)$ be the i^{th} child node. Consider U_{-n} a matrix obtained from U by discarding the n^{th} row. Let $\delta_{nm}^{(\ell)} = 1$ if $x_{nm}^{(\ell)}$ is non-missing and 0 otherwise.

Given the Markov blanket of $\mathbf{u}_n^{(\ell)}$ i.e., $\{\mathbf{x}_n^{(\ell)}, V^{(\ell)}, \mathbf{u}_{p(n)}^{(\ell-1)}, \mathbf{u}_{C(n)}^{(\ell+1)}\}$, $\mathbf{u}_n^{(\ell)}$ is independent of the other variables (Figure 6.1(a)). Therefore, the conditional probability of $U^{(\ell)}$ can be factorized into the product of conditional probability of its rows $\{\mathbf{u}_n^{(\ell)}\}_{n=1}^{N^{(\ell)}}$. By

applying Bayes rule and given that the product of multiple Gaussian distributions is another Gaussian distribution, it can be shown that the conditional probability of $\mathbf{u}_n^{(\ell)}$ is a Gaussian distribution

$$p\left(\mathbf{u}_n^{(\ell)}\middle|\mathbf{x}_n^{(\ell)}, V^{(\ell)}, \mathbf{u}_{p(n)}^{(\ell-1)}, \mathbf{u}_{C(n)}^{(\ell+1)}\right) \sim \mathcal{N}\left(\mathbf{u}_n^{(\ell)}\middle|\mu_n^{*(\ell)}, \Sigma_n^{*(\ell)}\right)$$
$$\sim \prod_m \left[\mathcal{N}\left(x_{nm}^{(\ell)}\middle|\langle \mathbf{u}_n^{(\ell)}, \mathbf{v}_m^{(\ell)}\rangle, \sigma^2\right)\right]^{\delta_{nm}^{(\ell)}} \prod_i \left[\mathcal{N}\left(\mathbf{u}_{c_i(n)}^{(\ell+1)}\middle|\mathbf{u}_n^{(\ell)}, \sigma_u^2 I\right)\right] \mathcal{N}\left(\mathbf{u}_n^{(\ell)}\middle|\mathbf{u}_{p(n)}^{(\ell-1)}, \sigma_u^2 I\right)$$

where $|.|$ denotes the set cardinality,

$$\Sigma_n^{*(\ell)} = \left[\frac{|C(n)|+1}{\sigma_u^2} I + \frac{1}{\sigma^2}\sum_m \delta_{nm}^{(\ell)} \mathbf{v}_m^{(\ell)} \mathbf{v}_m^{(\ell)T}\right]^{-1} \tag{6.1}$$
$$\mu_n^{*(\ell)} = \Sigma_n^{*(\ell)}\left[\frac{\mathbf{u}_{p(n)}^{(\ell-1)} + \sum_i \mathbf{u}_{c_i(n)}^{(\ell+1)}}{\sigma_u^2} + \frac{\sum_m \delta_{nm}^{(\ell)} x_{nm}^{(\ell)} \mathbf{v}_m^{(\ell)}}{\sigma^2}\right].$$

6.2.3 Sampling V

a) Block-wise Sampling: When $K_v^{(\ell)} = \frac{1}{\sigma_v} I$, each row of latent matrix $V^{(\ell)}$ can be sampled in parallel similar to sampling matrix $U^{(\ell)}$ (Section 6.2.2).
b) Element-wise Sampling: When $K_v^{(\ell)}$ is a full matrix, each column d of $V^{(\ell)}$ is drawn from $\mathcal{N}(V_{:d}^{(\ell)}|V_{:d}^{(\ell-1)}, [K_v^{(\ell)}]^{-1})$. Because of conditional dependencies, unlike sampling $U^{(\ell)}$ in Section 6.2.2, matrix $V^{(\ell)}$ is sampled element-wise.

By applying Bayes rule, the conditional probability of $v_{md}^{(\ell)}$ can be written as

$$p\left(v_{md}^{(\ell)}\middle|V_{-m,-d}^{(\ell)}, X^{(\ell)}, U^{(\ell)}, \mathbf{v}_{:d}^{(\ell-1)}, \mathbf{v}_{:d}^{(\ell+1)}\right) \sim \tag{6.2}$$
$$p\left(\mathbf{x}_{:m}^{(\ell)}|\mathbf{v}_m^{(\ell)}, U^{(\ell)}\right) p\left(v_{md}^{(\ell)}|\mathbf{v}_{-m,d}^{(\ell)}, \mathbf{v}_{:d}^{(\ell-1)}\right) p\left(v_{md}^{(\ell)}|\mathbf{v}_{-m,d}^{(\ell)}, \mathbf{v}_{:d}^{(\ell+1)}\right).$$

It can be shown that the individual distributions are univariate Gaussians as follows. Consider

$$p(\mathbf{x}_{:m}^{(\ell)}|\mathbf{v}_m^{(\ell)}, U^{(\ell)}) = \prod_n \mathcal{N}(\langle \mathbf{u}_n, \mathbf{v}_m\rangle, \sigma^2). \tag{6.3}$$

Given (6.3), the conditional probability of $v_{md}^{(\ell)}$ is obtained as

$$p\left(v_{md}^{(\ell)}|\mathbf{v}_{m,-d}^{(\ell)}, \mathbf{x}_{:m}^{(\ell)}, U^{(\ell)}\right) \sim \mathcal{N}\left(\mu_x^{(\ell)}, \frac{1}{\sigma_x^{(\ell)}}\right) \tag{6.4}$$

where $\sigma_x^{(\ell)} = \frac{\sum_n [u_{nd}^{(\ell)}]^2}{\sigma^2}$, $\mu_x^{(\ell)} = \frac{\sum_n u_{nd}^{(\ell)} \beta_n^{(\ell)}}{\sum_n [u_{nd}^{(\ell)}]^2}$, and $\beta_n^{(\ell)} = x_{nm}^{(\ell)} - \sum_{\substack{h=1 \\ h \neq d}}^{K} u_{nh}^{(\ell)} v_{mh}^{(\ell)}$.

Since the prior of each column $V^{(\ell)}$ is $\mathcal{N}(\mathbf{v}_{:d}^{(\ell)}|\mathbf{v}_{:d}^{(\ell-1)}, [K_v^{(\ell)}]^{-1})$, the conditional probabilities of $v_{md}^{(\ell)}$ can be obtained as follows,

$$\begin{aligned} p\left(v_{md}^{(\ell)}\middle|\mathbf{v}_{-m,d}^{(\ell)}, \mathbf{v}_{:d}^{(\ell-1)}\right) &\sim \mathcal{N}\left(\mu_{m1}^{(\ell)}, \frac{1}{\sigma_m}\right) \\ p\left(v_{md}^{(\ell)}\middle|\mathbf{v}_{-m,d}^{(\ell)}, \mathbf{v}_{:d}^{(\ell+1)}\right) &\sim \mathcal{N}\left(\mu_{m2}^{(\ell)}, \frac{1}{\sigma_m}\right) \end{aligned} \tag{6.5}$$

where $\sigma_m = K_v^{(\ell)}(m,m)$,

$$\begin{aligned} \mu_{m1}^{(\ell)} &= v_{md}^{(\ell-1)} - \frac{K_v^{(\ell)}(m,-m)}{K_v^{(\ell)}(m,m)} \left[\mathbf{v}_{-m,d}^{(\ell)} - \mathbf{v}_{-m,d}^{(\ell-1)}\right], \\ \mu_{m2}^{(\ell)} &= v_{md}^{(\ell+1)} - \frac{K_v^{(\ell)}(m,-m)}{K_v^{(\ell)}(m,m)} \left[\mathbf{v}_{-m,d}^{(\ell)} - \mathbf{v}_{-m,d}^{(\ell+1)}\right]. \end{aligned}$$

From (6.4), (6.5), we can write (6.2) as

$$p\left(v_{md}^{(\ell)}|V_{-m,-d}^{(\ell)}, X^{(\ell)}, U^{(\ell)}\mathbf{v}_{:d}^{(\ell-1)}, \mathbf{v}_{:d}^{(\ell+1)}\right) \sim \mathcal{N}\left(\mu_{md}^{*(\ell)}, \sigma_{md}^{*(\ell)}\right) \tag{6.6}$$

where $\sigma_{md}^{*(\ell)} = (2\sigma_m + \sigma_x^{(\ell)})^{-1}$,
$\mu_{md}^{*(\ell)} = \sigma_{md}^{*(\ell)} \left[\sigma_x^{(\ell)} \mu_x^{(\ell)} + \sigma_m(\mu_{m1}^{(\ell)} + \mu_{m2}^{(\ell)})\right]$.

6.2.4 BHPMF Inference

We consider three different sampling procedures based on selection of $K_v^{(\ell)}$ at each level as follows.

a) Block-wise Sampler: For a given sparse matrix X, the sampler updates the latent factor matrices $(U^{(\ell)}, V^{(\ell)})$ at every level ℓ. At each level ℓ, $U^{(\ell)}$ is sampled block-wise using (6.1). Using $K_v^{(\ell)} = \frac{1}{\sigma_v} I$ for all level $\ell = 1 \cdots L$, $V^{(\ell)}$ is sampled block-wise. To

incorporate the taxonomic information we use the following procedure. Each sample at the lowest level is obtained by sampling the upper level matrices iteratively. At each iteration, we first do a bottom-up pass to sample $(U^{(L)}, V^{(L)})$ to $(U^{(1)}, V^{(1)})$, followed by a top-down pass to sample $(U^{(1)}, V^{(1)})$ to $(U^{(L)}, V^{(L)})$, and repeat the procedure to generate enough samples (Algorithm 4).

b) Element-wise Sampler: At each level ℓ, similar to the block-wise sampler, $U^{(\ell)}$ is sampled using (6.1). To incorporate trait correlations into the sampler, a full covariance matrix $K_v^{(\ell)}$ is used for $\ell = 1 \cdots L$. Therefore, the matrix $V^{(\ell)}$ is sampled element-wise. The sampling procedure is mostly similar to the block-wise sampler (Algorithm 4) except that line 6 in Algorithm 4 is replaced with the following lines

---

6a: **for** $iter = 1 \cdots MaxIteration$ **do**
6b: **for** $m = 1 \cdots M$ **do**
6c: **for** $d = 1 \cdots D$ **do**
Sample $v_{md}^{(\ell)}$ using (6.6):
$p\left(v_{md}^{t+1(\ell)} | V_{-m,-d}^{t+1(\ell)}, X^{(\ell)}, U^{t+1(\ell)} \mathbf{v}_{:d}^{t(\ell-1)}, \mathbf{v}_{:d}^{t(\ell+1)}\right)$

---

where $MaxIteration$ is chosen empirically. Updating $V^{(\ell)}$ more than once at each iteration obtains a stable matrix before updating upper level matrices. Similar changes are applied to line 9 in Algorithm 4.

c) Mixture Sampler: At each level ℓ, similar to the block-wise sampler, $U^{(\ell)}$ is sampled using (6.1). For $\ell = 1 \cdots (L-1)$, $K_v^{(\ell)} = \frac{1}{\sigma_v} I$ is used and $V^{(\ell)}$ is sampled block-wised. At the lowest level L (plant level), a full covariance matrix $K_v^{(\ell)}$ is used and $V^{(\ell)}$ is sampled element-wise from (6.6).

6.3 Experimental Results

Here, we present the results for trait prediction.

6.3.1 Dataset

In our experiment, we use a cleaned subset of the TRY database – the world's largest database of plant trait [101]– where taxonomic hierarchy information is available for

Algorithm 4 BHPMF - Block-wise Sampler

for $\ell = 1, \cdots, L$ **do**
  Initialize model parameters $\{U^{1(\ell)}, V^{1(\ell)}\}$
for $t = 1, \cdots, T$ **do**
  for $\ell = L, \cdots, 1$ **do** ▷ bottom-up
    for $n = 1 \cdots N$ sample $\mathbf{u}_n^{(\ell)}$ in parallel using (6.1):
    $\mathbf{u}_n^{t+1(\ell)} \sim p\left(\mathbf{u}_n^{t(\ell)} | \mathbf{x}_n^{(\ell)}, V^{t(\ell)}, \mathbf{u}_{p(n)}^{t(\ell-1)}, \mathbf{u}_{C(n)}^{t(\ell+1)}\right)$
    for $m = 1 \cdots M$ sample $\mathbf{v}_m^{(\ell)}$ in parallel:
    $\mathbf{v}_m^{t+1(\ell)} \sim p(\mathbf{v}_m^{t(\ell)} | \mathbf{x}_m^{(\ell)}, U^{t+1(\ell)}, \mathbf{v}_m^{t(\ell-1)}, \mathbf{v}_m^{t(\ell+1)})$
  for $\ell = 1, \cdots, L$ **do** ▷ top-down
    for $n = 1 \cdots N$ sample $\mathbf{u}_n^{(\ell)}$ in parallel using (6.1):
    $\mathbf{u}_n^{t+2(\ell)} \sim p(\mathbf{u}_n^{t+1(\ell)} | \mathbf{x}_n^{(\ell)}, V^{t+1(\ell)}, \mathbf{u}_{p(n)}^{t+1(\ell-1)}, \mathbf{u}_{C(n)}^{t+1(\ell+1)})$
    for $m = 1 \cdots M$ sample $\mathbf{v}_m^{(\ell)}$ in parallel:
    $\mathbf{v}_m^{t+2(\ell)} \sim p(\mathbf{v}_m^{t+1(\ell)} | \mathbf{x}_m^{(\ell)}, U^{t+2(\ell)}, \mathbf{v}_m^{t+1(\ell-1)}, \mathbf{v}_m^{t+1(\ell+1)})$

all entries. This subset is a matrix containing 78,300 plants and 13 traits Table 6.1). The percentage of missing entries varies from 49.63% to 92.33% for each trait. In total, 79.9% of entries are missing. Starting from the top of the taxonomic hierarchy, there are 6 phylogenetic groups, 358 families, 3793 genera, 14,320 species, and 78,300 plants.

It has been discovered that plant traits are characterized by log-normal distributions [101]. We transformed the *plant*×*trait* matrix by taking the logarithm of entries followed by the z-score of traits. After this step, the trait values ranged from -4 to 4. The results we show are in the transformed space.

Given the $plant \times trait$ matrix and the taxonomic hierarchy, trait data matrices at upper levels, such as $species \times trait$ matrix, $genus \times trait$ matrix, etc. are constructed. For example, a $species \times traits$ matrix can be constructed by taking the average of the plants in the same species.

6.3.2 Baselines

Mean: Given the $plant \times trait$ training matrix, upper level matrices are constructed to provide species mean, genus mean, etc. using taxonomic information. For example, species mean of trait m is the average of trait m among plants in the same species with available trait m. To predict missing trait m of plant n, among species mean, genus

Table 6.1: ID, name, percentage of missing entries (%) and definition of the respective trait.

ID	Trait	Missing	Definition
1	Specific leaf area (SLA)	57.85	One sided area of a fresh leaf divided by its oven-dry mass
2	Plant height	78.97	Shortest distance of photosynthetic tissue or reproduction unit on a plant and the ground level
3	Seed dry mass	90.66	Dry mass of a whole single seed
4	Leaf dry matter content (LDMC)	77.87	Leaf dry mass per unit of leaf fresh mass (hydrated)
5	Stem specific density	88.26	Oven-dry mass of a section of a plant's main stem divided by its fresh volume
6	Leaf area	49.63	One-sided projected surface area of a single leaf or leaf lamina
7	Leaf nitrogen concentration (LeafN)	65.67	Total amount of nitrogen per unit of leaf dry mass
8	Leaf phosphorus concentration (LeafP)	84.71	Total amount of phosphorus per unit of leaf dry mass
9	Leaf nitrogen per area	89.55	Total amount of nitrogen per unit of leaf area (one-sided)
10	Leaf fresh mass	85.33	Fresh mass of a whole leaf
11	Leaf nitrogen/phosphorus ratio	92.34	Ratio of leaf total nitrogen versus total phosphorus
12	Leaf carbon per dry mass	89.63	Total carbon per unit of leaf dry mass
13	Leaf δ^{15}N	88.48	Foliar ^{15}N:^{14}N ratios relative to ^{15}N:^{14}N ratios in atmospheric N_2

mean, etc. we use the first available one at the lowest level.

PMF: We run PMF [174] on *plant* $\times$ *trait* matrix directly. Note that PMF is unable to consider the taxonomic information.

HPMF: The results of HPMF are obtained from 5 top-down and bottom-up passes in total same as [183].

6.3.3 Methodology

The data are split into training, validation, and test set similar to [183] as follows. We use plants with only one available trait in the training set. If a plant has two available traits, we keep randomly one trait for training and the other trait for test. If there are more than two traits available, we keep randomly one for test, one for validation, and the rest for training. The above holding out procedure is done 5 times. Note that, for upper level matrices only training and validation sets are constructed.

The most common evaluation measure for prediction accuracy is the root of the mean square error (RMSE), given as $RMSE = \frac{1}{T}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{ij}(x_{ij}-\hat{x}_{ij})^2}$ where x_{ij} is the actual trait value, $\hat{x}_{ij}$ is the predicted value for plant i and trait j, and T is

Table 6.2: RMSE of Species Mean, PMF, HPMF and BHPMF. Latent dimension k=15 for matrix factorization methods.

Method	*RMSE*
PMF	0.8993 ± 0.0210
MEAN	0.5753 ± 0.0024
HPMF	0.5009 ± 0.0034
BHPMF - Block-wise Sampler	$\mathbf{0.4567 \pm 0.0021}$

the total number of non-zero entries .

For uncertainty evaluation, we report our results based on a model's confidence vs. accuracy curve. We use the standard deviation to measure the degree of confidence in trait prediction, and RMSE to measure the model's accuracy. The hypothesis is that when we are confident in the predictions on the test set, the achieved accuracy is high i.e., the standard deviation should decrease with decreasing RMSE.

In order to run BHPMF, the latent matrices $U^{(\ell)}$ and $V^{(\ell)}$ for $\ell = 1 \cdots L$ are initialized randomly. The parameters for different BHPMF samplers are as follows.

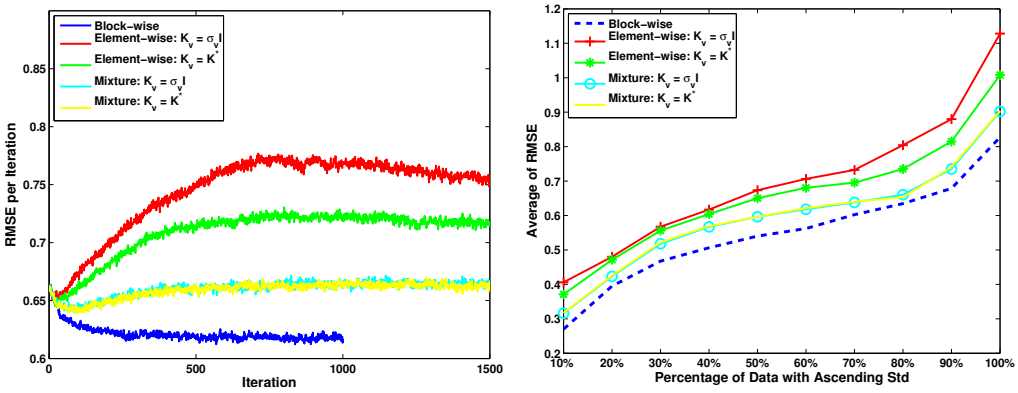
Block-wise sampler: The burn-in period was set to 200 with a lag of 2 and final number of samples 400.

Element-wise sampler: The burn-in period was set to 700 with a lag of 2 and final number of samples 400. Element-wise sampler has been tested with $K_v^{(\ell)} = \frac{1}{\sigma_v} I$ and $K_v^{(\ell)} = K^*$ where K^* is the estimated precision matrix by mGLasso (an estimator of precision matrix with missing value) [103].

Mixture sampler: The burn-in period was set to 700 samples with a lag of 2. The final number was samples of 400. Mixture sampler has been tested with both $K_v^{(\ell)} = \frac{1}{\sigma_v} I$ and $K_v^{(\ell)} = K^*$.

6.3.4 Results

In this Section, we evaluate BHPMF in different aspects like comparison between different samplers , uncertainty evaluation analysis, and prediction accuracy.



(a) (b)

Figure 6.2: a) RMSE of different BHPMF samplers with increasing number of iterations. Block-wise sampler outperforms others. b) BHPMF for all traits and with the inverse of prediction confidence (Std) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the Std is small (more confident).

Different Type of Samplers

Figure 6.2(a) illustrates a comparison between different BHPMF samplers with respect to RMSE per iteration. All of the samplers reach to a stationary state with increasing number of iterations. Interestingly, the block-wise sampler with $\{K_v^{(\ell)} = \frac{1}{\sigma_v} I\}_{\ell=1}^{L}$, outperforms all other samplers. While mixture sampling with both type of covariance matrix behave almost similarly, the element-wise sampler with full covariance $\{K_v^{(\ell)}\}_{\ell=1}^{L}$ improves the element-wise sampler with a diagonal covairance matrix.

Uncertainty Evaluation

The experiment runs as follows: we sort all the data points in the test sets in ascending order of their standard deviation (Std), and divide the test sets evenly into 10 parts according to ascending Std, i.e., the first part (Batch 1) contains the first 10% data points with the lowest Std, the second part (Batch 2) contains the second 10% data points with the second lowest Std, and so on. We calculate the RMSE on these 10 parts separately and draw a curve. Figure 6.2(b) and 6.6 illustrates the curve on the 13-trait TRY data set. It is observed that the RMSE increases monotonically with increasing Std. By looking at the Std vs RMSE curve, we conclude that when we are confident

about our predictions, the predictions are accurate. Since higher RMSE indicates lower accuracy, and higher standard deviation indicates lower confidence, the observation could be rephrased as: the model's accuracy decreases monotonically with decreasing the model's confidence, i.e., the less confidence the model has, the worse performance it gets. Therefore, our hypothesis is verified. Uncertainty quantification not only is a tool to measure how accurate the predicted trait values are, but also provides the areas of less confident predictions which can in turn be used to guide field work for data collection efforts. Similar results have been observed with considering each trait separately.

Geographical Distribution of Uncertainties

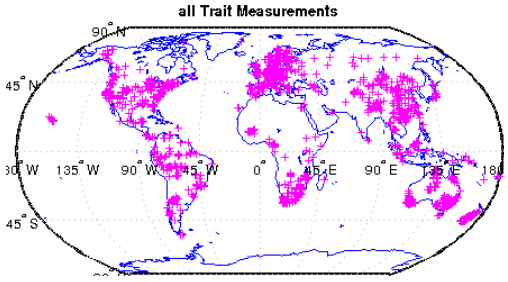
In order to identify areas of limited confidence, we explored the spatial coverage of different batches with different uncertainties (Figure 6.3). It can be observed that trait measurements are scattered in a wider range of the world by going to more uncertain batches i.e., going from batch 1 to batch 10. Particularly, trait measurements in China or south Africa have been appeared more in the uncertain batches. Additional measurements even in the densely covered regions like China or South Africa may improve the accuracy.

Prediction Accuracy

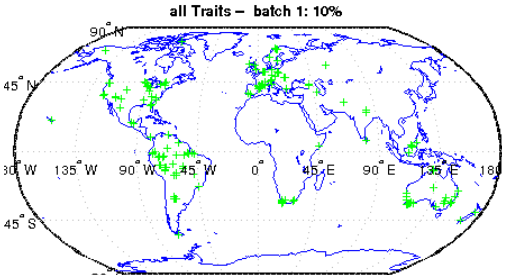
We also compared the point estimation derived from BHPMF with MEAN, PMF, and HPMF. As shown in Section 6.3.4, the block-wise sampler outperforms the other sampling types. Therefore, we only provide point estimation results of the block-wise sampler. BHMPF provides the point estimation of each missing trait value by taking the average of all generated samples. The RMSE results are shown in table 6.2. BHPMF outperforms all the other models, which means BHPMF not only provides uncertainty quantification but also improves the point estimation of current trait value predictions.

Trait Correlation

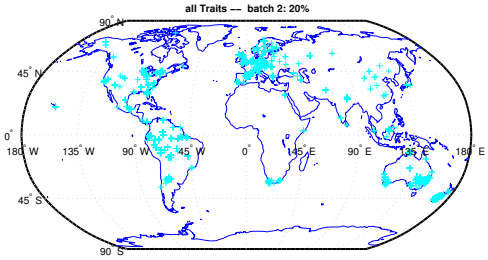
In ecological community, multivariate joint trait analyses are highly desired. Traits do not occur in isolation of each other, but rather in the form of plant individuals, e.g., small



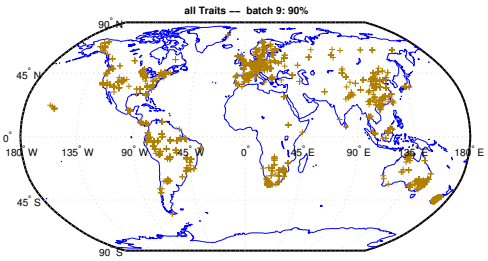
(a) All Observations



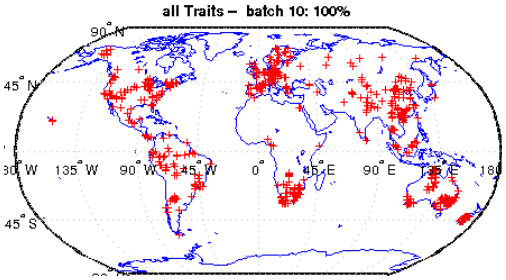
(b) The highest confident group



(c) (b) 2^{nd} highest confident group



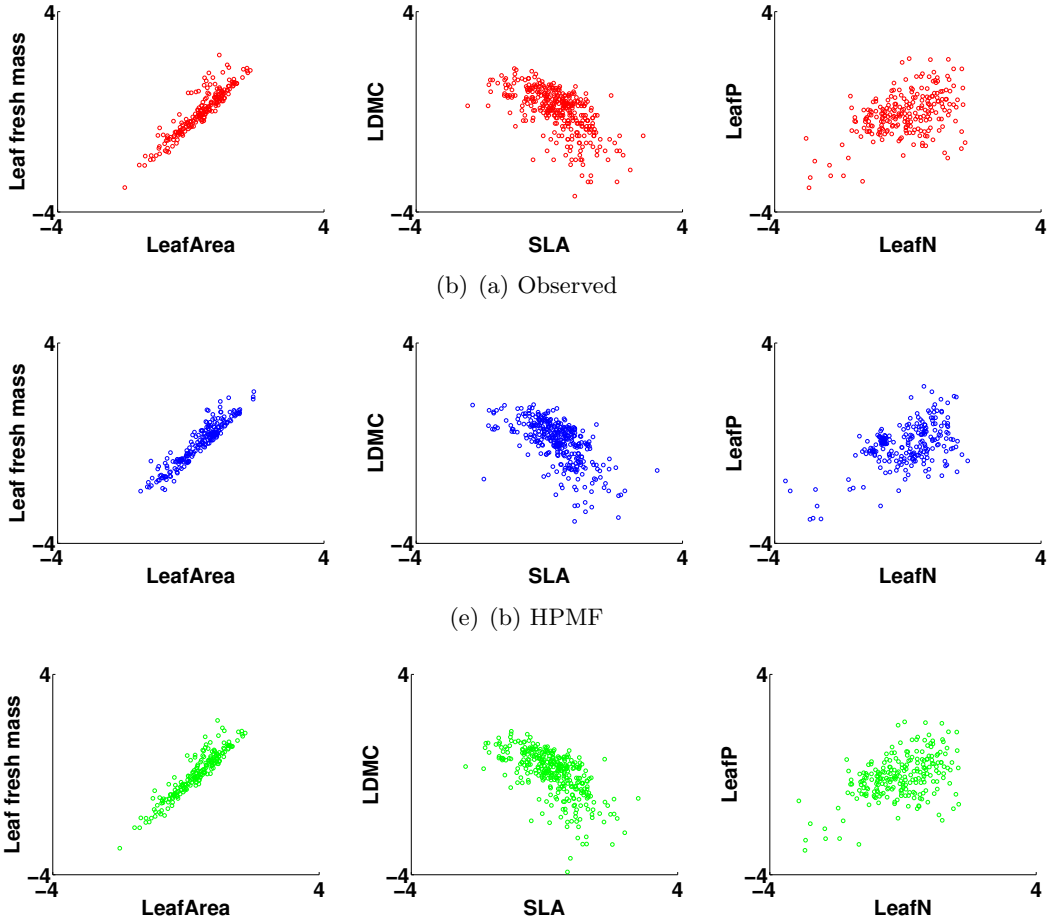
(d) (c) 2^{nd} lowest confident group



(e) The lowest confident group

Figure 6.3: a,b,e) Spatial coverage of all observation, the highest and lowest confident group. Trait measurements in China or south Africa are more frequent in the uncertain groups (e). Additional measurements in the densely covered regions like China may improve the accuracy.

plants cannot have big seeds; high photosynthesis rates depend on high concentrations of respective enzymes. This multidimensional correlation structure between different traits has led to the definition of trait syndromes [17, 44, 63, 224]. However, with a sparse matrix like TRY, this kind of analysis is challenging and impossible for plants with only one available trait. The relevance of triat-trait correlations motivates us to test matrix completion with another set of model evaluation, to check how close predicted trait



(b) (a) Observed

(e) (b) HPMF

(h) (c) BHPMF

Figure 6.4: Scatter plots for pairs of traits (a) on observed true test data, (b) predicted by HPMF, and (c) predicted by BHPMF. BHPMF and PHMF preserve true trait correlations.

correlations are to true correlations.

For plants with available pairs of traits in the test set, the predicted correlation can be compared with the observed true correlation by drawing the corresponding scatter plots. As an example, we provide the scatter plots for Leaf Fresh Mass vs. Leaf Area, SLA vs. LDMC, and LeafN vs. LeafP (Figure 6.4). It can be observed that both HPMF and BHPMF preserve the true correlation for all three pairs.

6.4 Multiple Inheritance BHPMF

The Multiple Inheritance model we present here can be addressed as a generalization of the BHPMF model. In the BHPMF model, $u_n^{(\ell)}$ and $v_m^{(\ell)}$ are generated from a single Gaussian distribution with a mean around their parents latent factors at the higher level. In the case of multiple inheritance, the BHPMF model could be generalized to generate each latent factor $u_n^{(\ell)}$ and $v_m^{(\ell)}$ from product of Gaussian distributions, involving a subset of parents. Markov blanket of Multiple Inheritance BHPMF (MI-BHPMF) is illustrated in Figure 6.1(b). MI-HPMF assumes a directed acyclic graph (DAG) structured hierarchical prior, rather than tree structured as in BHPMF. The construction has parallels with the product of expert models [89], and multiplicative mixture models (MMM) [79, 86]. A key difference here is that the DAG structure is assumed to be known, and here a combination of inference in MMM [79, 86] is avoided.

The generalized model of MI-BHPMF at level ℓ is:

1. Generate $\mathbf{u}_n^{(\ell)} \sim \prod_i N(\mathbf{u}_{p_i(n)}^{(\ell-1)}, \sigma_u^2 I)$, $[n]_1^{N^{(\ell)}}$.
2. Generate $\mathbf{v}_{:d}^{(\ell)} \sim N(\mathbf{v}_{:d}^{(\ell-1)}, [K_v^{(\ell)}]^{-1})$, $[d]_1^D$.
3. Generate $x_{nm}^{(\ell)} \sim N(\langle \mathbf{u}_n^{(\ell)}, \mathbf{v}_m^{(\ell)} \rangle, \sigma^2)$ for each non-missing entry.

where $p_i(n)$ is the i^{th} parent of n in the upper level.

In principle, V can also have multiple inheritance. Here, we discuss the case where only U has multiple inheritance. The conditional probability of $\mathbf{u}_n^{(\ell)}$ is

$$p\left(\mathbf{u}_n^{(\ell)} \middle| \mathbf{x}_n^{(\ell)}, V^{(\ell)}, \mathbf{u}_{P(n)}^{(\ell-1)}, \mathbf{u}_{C(n)}^{(\ell+1)}\right) \sim \mathcal{N}\left(\mathbf{u}_n^{(\ell)} \middle| \mu_n^{*(\ell)}, \Sigma_n^{*(\ell)}\right)$$

where $P(n) = \{p_i(n)\}$ is the set of parent nodes of n,

$$\Sigma_n^{*(\ell)} = \left[\frac{|C(n)| + |P(n)|}{\sigma_u^2} I + \frac{\sum_m \delta_{nm}^{(\ell)} \mathbf{v}_m^{(\ell)} \mathbf{v}_m^{(\ell)T}}{\sigma^2}\right]^{-1} \tag{6.7}$$

$$\mu_n^{*(\ell)} = \Sigma_n^{*(\ell)} \left[\frac{\sum_i \mathbf{u}_{p_i(n)}^{(\ell-1)} + \sum_j \mathbf{u}_{c_j(n)}^{(\ell+1)}}{\sigma_u^2} + \frac{\sum_m \delta_{nm}^{(\ell)} x_{nm}^{(\ell)} \mathbf{v}_m^{(\ell)}}{\sigma^2}\right].$$

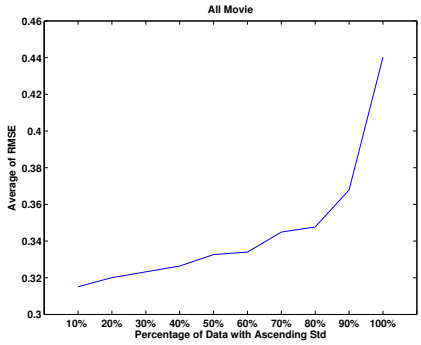


Figure 6.5: MI-BHPMF for all movies with the inverse of prediction confidence (standard deviation) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the standard deviation is small (more confident).

The sampling procedure is mostly similar to Algorithm 4 except that line 5 is replaced with the following line

---

5: **for** $n = 1 \cdots N$ sample $\mathbf{u}_n^{(\ell)}$ in parallel using (6.7):
$\mathbf{u}_n^{t+1(\ell)} \sim p\left(\mathbf{u}_n^{t(\ell)} \middle| \mathbf{x}_n^{(\ell)}, V^{t(\ell)}, \mathbf{u}_{P(n)}^{t(\ell-1)}, \mathbf{u}_{C(n)}^{t(\ell+1)}\right)$

Similar change is applied to line 8 in Algorithm 4 for the top-down procedure.

We present some preliminary results of evaluating the multiple inheritance model on the MovieLens Data set. The data set contains 1M ratings for 3900 movies by 6040 users. The genre of each movie has been extracted from IMDB [182]. There are 25 movie types (Genre). A hierarchy over movies can be built by grouping movies based on genre where each movie may belong to more than one genre. Figure 6.5 shows RMSE-Std curve on the MovieLens data set. Similar to BHPMF, RMSE increases monotonically with increasing standard deviation. Meaning that MI-BHPMF is accurate (small RMSE) when it is confident (small standard deviation).

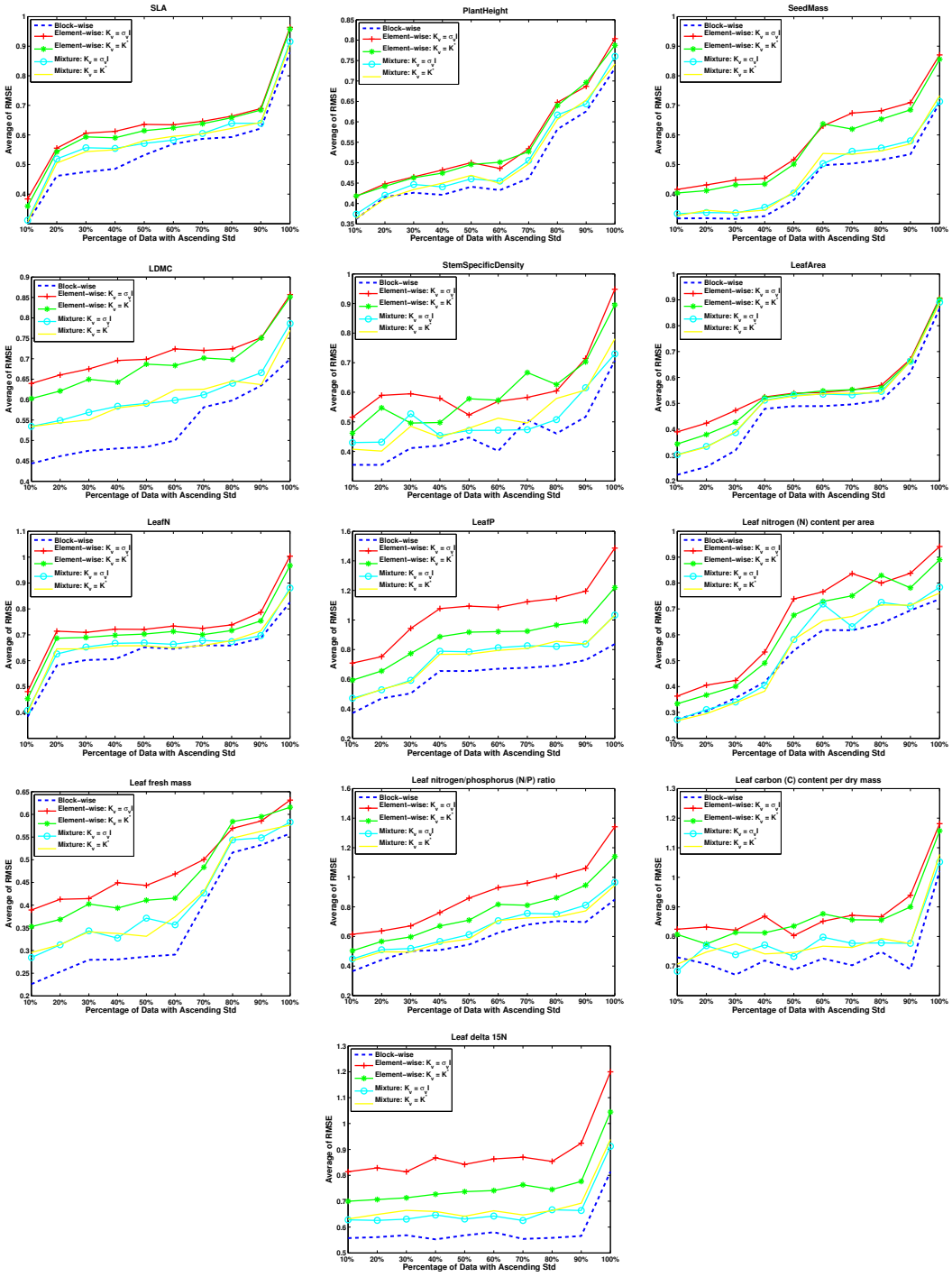


Figure 6.6: BHPMF for each of the 13 traits with the inverse of prediction confidence (standard deviation) on the x-axis and the prediction error (RMSE) on the y-axis. The errors are small (more accurate) when the standard deviation is small (more confident).

Part III

Application

Chapter 7

Trait-Trait Interactions across Climate Zones

7.1 Statement of Contribution of co-authors

This study is a joint work with Habacuc Flores-Moreno, Arindam Banerjee, Abhirup Datta, Jens Kattge, Ethan E. Butler, Owen K. Atkin, Kirk Wythers, Ming Chen, Madhur Anand, Michael Bahn, Sabina Burrascano, Chaeho Byun, J. Hans C. Cornelissen, Joseph Craine, Andres Gonzalez-Melo, Wesley N. Hattingh, Steven Jansen, Nathan J.B. Kraft, Koen Kramer, Daniel C. Laughlin, Vanessa Minden, Ülo Niinemets, Vladimir Onipchenko, Josep Peñuelas, Nadejda A. Soudzilovskaia and Peter B. Reich [73].

HFM, FF, AB, PBR designed the study. FF and AB developed statistical method. JK, HFM and FF prepared the data. HFM and FF analyzed data. HFM and FF wrote the manuscript, with all authors contributing to subsequent revisions.

7.2 Introduction

Because plant traits are not independent of each other biologically or statistically, an accurate description of their interdependency gives us a clearer view of the links morphological traits and physiological function [156, 161, 219], differences between functional groups [155], the effect of multivariate trait relationships on mechanisms of coexistence [108] and ecosystem processes modeled at global scales [215]. However, the strength and

form of the relationships between traits among co-occurring species varies across environments [170, 3], and among functional groups [169], with axes of trait variation shifting, collapsing or arising across environmental gradients and different plant life forms. Empirical support for generalities in the coordination of traits exist, with evidence mostly limited to single organs or few traits at broad spatial scales [224, 45, 148, 237] or across organs but only for certain regions and/or growth forms [3, 18, 75, 74, 50, 109]. Paving the way forward, Diaz et al. [64] analyzed the major axes of trait variation across the plant kingdom for six traits measured on different organs, finding strong evidence for coordination (i.e. non-random variation between traits) among these traits at global scales. As of yet, however, it remains unknown whether the connections among multiple traits across organs within environments are similar among species and across environments. We use connection in a specific way in this manuscript; as representing direct functional linkages between traits, which we evaluate through a test of statistical independence that accounts for co-variation with all other traits in the data set. This approach thus allows us to address (i) whether correlations among traits all represent connections or whether some are solely due to co-variance with other traits, and (ii) additionally whether weakness in simple bivariate correlations might mask connections. We then assess whether and how such connections collectively describe networks across scales of broad climate regions and plant life forms.

Previous studies that focused on multi-organ, multi-trait datasets have typically been limited in geographic scope and described the main axes of trait variation, the association of traits to each of these axes, and associated trait correlations and trade-offs [3, 75, 109]. Usually, two axes are defined and have been interpreted as a resource uptake axis, and a second axis related to competitive ability, and/or response to disturbance [223, 18]. However, up to four independent axes may be described [96]. Evidence suggests that the trait composition of these axes, whether interpreted as niche dimensions or organ axes, changes across communities under different environmental conditions. For instance, the independence of root, leaf, and stem traits varies across environmental gradients, with plant communities from wetter environments showing the leaf axis to be orthogonal to the root and stem axes [18, 74, 96], while in drier environments orthogonality across these organs is lost [115, 121, 62]. Similarly, the independence of traits related to water economy from the resource acquisition axis increases in environments

with higher precipitation[3, 94, 18, 74], a pattern also observed across a soil water gradient within Mediterranean vegetation[62]. In tundra, phosphorus (P) content weakly contributes to the resource uptake axis [75], while studies from warmer and wetter regions (shrublands, temperate forests and tropical forest) show P content to be tightly linked to the resource uptake axis and plant nitrogen economy [18, 74, 96, 121, 109]. Given differences among floras in the number of main axes of trait variation, the trait composition of these axes, and independence among organs, it is still unclear whether the integration (i.e. level of cross-linked trait connection across the plant) of traits changes in a predictable way across broad environmental gradients, or whether multiple strategies for different environmental conditions or plant types exist.

Investigating the generality of the relationship between suites of traits within and among floras has many challenges as understanding these relationships requires considering many species and many traits ([3]; e.g. Table 7.1). At this stage, and despite the increased number of studies of multi-trait datasets from diverse floras, it is still challenging to separate general patterns in the integration of plant strategies in response to environmental gradients, from idiosyncrasies of the vegetation types or study. This is partly due to across-study comparisons being complicated by differences in: the number/type of traits used (Table 7.1); the ontogenetic stage of the plants measured (e.g. [94, 121, 109]); differences in the environmental conditions of the sampled plants (e.g. [223, 121, 109]); and/or differences in operational definition of traits (e.g. fine roots; [50, 121, 62]). Also, spatial scales tend to vary widely (see Table 7.1) and how differences in scale affect comparisons across studies focused on multi-trait datasets has yet to be explored. A study on leaf trait variation and integration did, however, suggest that biological and spatial scale could have important implications for trait integration [142]. Nuances across studies are many in terms of the number and kinds of trait metrics used, the relative importance of traits to the plant strategies dimensions or the degree of connectivity between traits/organs across environmental gradients.

An equally important challenge in understanding connections among traits is interpreting evidence from correlative analyses and dimension reduction techniques. For instance, commonly used ordination techniques such as principal component analyses (PCA) effectively decrease the dimensionality of multivariate trait datasets and are easily combined with standard statistical methods (e.g. linear models). However, the

biological interpretation of the PC axes and of the traits composing an axis requires a certain level of intuition, and may become fuzzy, especially as the number of axes considered increases [160, 119]. Further, standard practices such as only selecting a subset of PC axes prevents a full exploration of the multivariate nature of plant strategies and can have unintended consequences (e.g. erroneous inference due to bias sample of multivariate distribution; see [207]). On the other hand, while correlative analyses generally capture the marginal dependency among variables [122, 119], it is impossible to detect indirect from direct dependencies among variables from correlation analyses alone [218, 185, 37, 67, 156]. Distinguishing direct and indirect connections among traits is necessary for understanding the mechanistic roots of those trait correlations that define plant strategies, and can thus help us clarify the causal link between traits and fitness components [185], connections among traits and function [161], and the role that traits play in influencing higher-level processes and vegetation attributes (e.g. RGR, NPP, [167]).

In this study, we describe the connections among traits across organs in the uncollapsed trait multidimensional space using precision matrices (inverse of covariance matrix) a special case of probabilistic graphical models [67]. The key feature of this method is its capacity to identify direct from indirect connections by describing relationships among traits once the influence of the trait constellation has been considered. Then, differences in the trait connections across groups of plants (e.g. herbs vs. shrubs, monocots vs. dicots) can be compared based on the graph topology derived from the trait precision matrices for those groups. We use network theory to interpret the graph topology of the trait network obtained from the precision matrices (see Methods).

We focus on the following, overarching question: how do connections among traits vary across growth forms (woody and non-woody) and environmental gradients? Strong integration across traits across different organs is expected as matching tissue strategies should be advantageous at the whole-plant scale [168]. However, whether coordination across traits across organs is advantageous or not will depend on the local environmental factors that plants experience, as in any given environment not all resources necessary for a plant to persist and grow may be available in the same relative supply at the same time [24]. Further, woody species have a long-persisting, reinforced stem aboveground that allow vertical and lateral expansion (via cambium inside trunks or

broadly branching canopies), while non-woody species lack such a stem. This is a crucial difference across terrestrial plants, with profound impacts in the phenology, reserve patterns, and biophysical requirements of these groups and may influence the degree of integration among organs and their traits [24], which may be reflected in differences in the connectivity among traits of woody and non-woody species.

In summary, in the current study we: (i) describe the global trait network (statistical dependency between ten functional traits) among 16,281 plant species from sites around the world; (ii) compare the differences in the trait networks of woody and non-woody species; and (iii) assess how the trait network of woody and non-woody species changes across five broad climate regions (tropical, temperate, arid, cold and polar).

7.3 Method

7.3.1 Data

Our attention is on ten traits relevant to resource economy and uptake, competitive ability (or stress tolerance) and reproductive strategy of plants. Seed mass (mg) reflects allocation of energy to a few large vs. many small offspring, and impacts early seedling survival [149]. Plant height (m) and stem specific density (mg dry mass mm-3 fresh volume; hereon SSD) are traits related to light competition, growth rate and long-term viability of the stem [45, 148, 65]. Specific leaf area (mm2 mg-1; SLA), leaf lifespan (LLS; month), leaf nitrogen (N; mg g-1) mass, leaf phosphorus (P; mg g-1) mass are traits related to nutrient economy and acquisition and are key components of the leaf economic spectrum (LES; [170, 224]. SLA represents the mass investment related to a potential rate of return measured in terms of light capture area [171, 224]. Leaf lifespan represents the time needed to generate payback on this mass investment [167]. Leaf N is associated with carboxylation-capacity and is integral to the photosynthetic machinery [170, 224]. Leaf P is essential for bioenergetics molecules (e.g. ATP) and is indispensable for the formation of nucleic acids and lipid membranes [170, 224]. Therefore variation in leaf P and N will be crucial to respiration and photosynthetic capacity of plants, as well as energy generation and storage. Leaf area (mm^2) is related to the water and energy balance of a plant and is relevant to light interception. Finally, leaf N and P can also be expressed on an area basis reflecting light capture and transaction of energy on an

area basis (g m^{-2}) reflecting light capture and transaction of energy on an area basis [224]. Consequently, we also use leaf N per area and leaf P per area, in parallel to their mass-based counterparts.

We obtained spatially explicit trait data for our ten traits and growth form (woody, non-woody) data from TRY(www.try.db; [101], Reference for individual studies included in this database are provided in Appendix S1). The TRY data subset used in this study includes 19725 records for 16281 species across all terrestrial biomes sensu [221], from which 9053 and 6231 records were identified as either woody or non-woody plants respectively. We standardized the species names and higher order taxonomy according to the plant list (The Plant List, 2013) using Taxonstand [40], then we obtained the higher order taxonomy (i.e. family, order, group) for our species with taxonlookup (version 1.0.1; [159]). The number of individual records in global datasets rapidly decreases, as information on more traits is required [101]. In our dataset LLS was the trait with the lowest number of records (0.67% of records have information for this trait), while plant height was the trait with the highest number of records present (35%). This limits our ability to assess how traits vary jointly. Thus, we used a hierarchical Bayesian extension of probabilistic matrix factorization to fill in the trait gaps in our dataset [71, 179, 183]. This algorithm harnesses the available trait information and the species and higher order taxonomy to fill in the gaps in the trait data. This gap-filling method has been used in other trait analysis at global scale with robust results [64]. We also checked the robustness of this method by comparing the trait-trait relationship of only gap-filled data vs. only original data (Table 7.2).

Using a map of the Köppen climate zones we assigned the georeferenced plant records to five different climate zones: tropical climate, which includes tropical rainforest, tropical seasonal forest, and savannahs; arid climate, which includes deserts and steppes; temperate climate, which includes temperate forest, temperate rainforest, and Mediterranean vegetation; cold climate, which includes only taiga; polar climate, which includes tundra, alpine and circumpolar zones (see [158]). The climate regions described above are derived from a combination of global patterns in temperature and precipitation ([158]). At the same time, important differences in chemical and anatomical traits exist between woody angiosperms vs. gymnosperms and non-woody forbs vs. monocots [56, 203, 26, 45]). Thus, to distinguish the effect of precipitation and temperature, as

well as that of differences between non-woody forbs vs. monocots and woody gymnosperms vs. angiosperms we compared the connection among traits for these groups of species across a precipitation gradient, holding temperature constant, and a temperature gradient holding precipitation constant (Section 7.4.6; Tables 7.11, 7.10; Figures 7.7, 7.9).

7.3.2 Analysis

We applied our new method using precision matrices to determine trait connections across multiple traits across organs. Given a constellation of traits, the precision matrix establishes the conditional independence among traits. When a coefficient in the precision matrix of traits is different from zero, then a relationship between traits x and y are not due to variation in z (z being a single trait or a set of traits; [92]; Figure 7.1b). On the other hand, a value of zero reveals the conditional independency between two traits, given variation in other traits (Figure 7.1c). For instance, if trait x and y are conditionally independent given trait z, that means that traits x and y do not provide information about each other once trait z is considered (Section 7.4.7). From this it follows that even if traits x and y are highly correlated, once trait z is considered any direct relationship between traits x and y would disappear. Thus, the precision matrix of traits provides the statistical conditional dependency structure among traits for a multivariate (in this case, assumed Gaussian) trait constellation, which is a graph structure that describes direct probabilistic interactions among traits (Section 7.4.7; [37, 67]).

We used a 'glasso' algorithm (Graphical Lasso; [77]) to estimate the precision matrix for each plant group with confidence intervals for each trait-trait interaction in the trait network. The glasso algorithm assumes that traits have a multivariate Gaussian distribution and estimates the precision matrix by maximizing the log-likelihood among all plant trait measurements (Notes S2). The glasso algorithm includes a penalty parameter λ, which controls the sparsity level of the precision matrix. Following [98], we accounted for differences in sample size across precision matrices in the calculation of λ as follows: $\lambda = 2\sqrt{\frac{logp}{n}}$, where n refers to the sample size and p denotes the number of nodes, in this case traits. First, we derived the trait network for all plants, and then for non-woody and woody plants separately. Next, we derived the trait network for

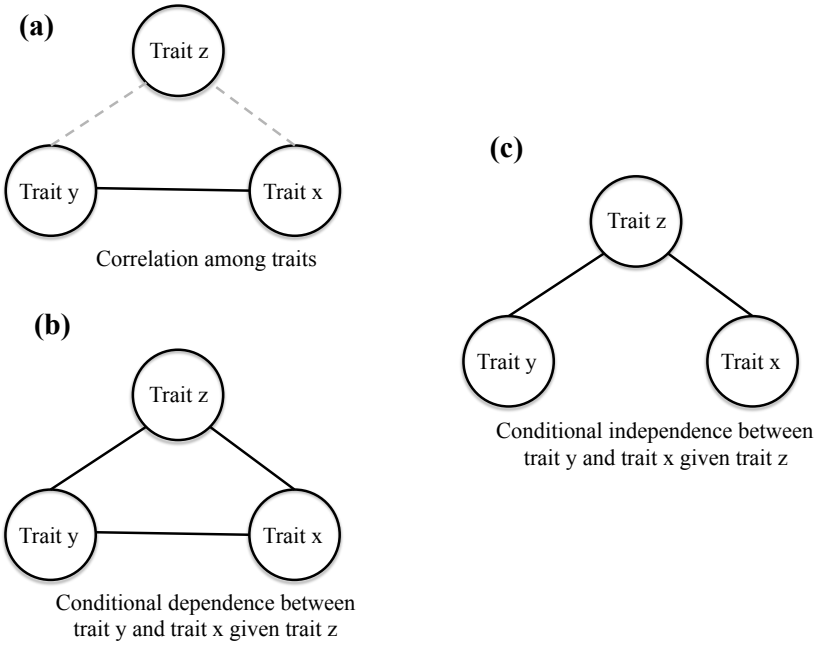


Figure 7.1: Hypothetical scenarios when determining the conditional dependency among correlated traits. (a) Observed correlation between trait y and trait x, when the effect of trait z has not been considered (dashed gray lines). (b) Conditional dependence between trait x and trait y even after considering trait z, suggesting dependency between trait x and trait y. (c) Conditional independence between trait y and trait x, once trait z has been considered, suggesting that the correlation between x and y was indirectly mediated through z.

non-woody and woody plants across the five climate zones defined above.

At the end, we test the significance of obtained edges i.e., trait-trait interactions, for each group by following [98]. Given a data sample matrix $X \in \mathcal{R}^{n \times p}$ and estimated precision matrix $\Omega \in \mathcal{R}^{p \times p}$, define the test statistics $T = \Omega + \lambda \Omega Z \Omega$ where Z is the sub-gradient of norm $\|\Omega\|_1$ and $\hat{\sigma}_{ij}^2 = \Omega_{ii}\Omega_{jj} + \Omega_{ij}^2$. Then, J. Jankova and S. van de Geer [98] show that thresholding T_{ij} at level $\Phi^{-1}\left(1 - \frac{\alpha}{p(p-1)}\right)\frac{\hat{\sigma}_{ij}}{\sqrt{n}}$ for all i, j will remove all zero entries (i.e., non-significant trait-trait interactions) with probability $1-\alpha$ asymptotically.

We estimate the $(1-\alpha)$ asymptotic confidence interval of all the obtained edges as

$$I_{ij} = \left[T_{ij} - \Phi^{-1}\left(1 - \frac{\alpha}{p(p-1)}\right)\frac{\hat{\sigma}_{ij}}{\sqrt{n}},\ T_{ij} + \Phi^{-1}\left(1 - \frac{\alpha}{p(p-1)}\right)\frac{\hat{\sigma}_{ij}}{\sqrt{n}} \right]. \tag{7.1}$$

We remove edges whose confidence interval contain zero.

To assess differences in the network topology of the different plant groups we used three network metrics:

Degree: ([92]; number of connections between a focal trait and other traits, normalized by the total number of unique connections) to quantify the relative importance of the traits for a given trait network. Degree is widely used in biological networks to identify essential characteristics in biological entities such as genes, metabolites, and proteins [107]. For example, evidence suggests that metabolites with higher Degree may belong to the oldest part of the metabolism, while proteins with higher Degree have been identified as essential, with their removal being lethal to the organism [107]. The values for Degree range between zero when a focal trait has no connection to any other trait, to one when a focal trait is connected to all other traits in the network.

'Modularity' is the difference between the fraction of connections among traits that fall within a given module (i.e. a module is a subset of traits, that interact more among themselves than with other surrounding traits) minus the same fraction in a null model were connections among traits are distributed randomly [54]. Thus, we used Modularity to measure how separated traits belonging to different modules are from each other in terms of fraction of connections that occur within modules (i.e. how connected/disconnected traits across modules are). Modularity helps identify nodes within a network that perform a common function and interact strongly among themselves. Higher Modularity confers an advantage under variable conditions as it provides robustness [163, 7], also providing opportunities for the network to adapt and evolve, as not all components in a modular network are optimally linked (i.e. ability to respond to changing external conditions/internal organization while maintaining normal behaviour; [16]. We used a Spinglass algorithm to detect the modules in our networks, as this algorithm accurately detects modules in networks of small size (Number of nodes ≤ 233) and with small or large mixing parameters at the network level (i.e. the summation of external

degree of each node over the summation of its total degree) [230]. In Modularity values of zero represent networks without a community structure, while non-zero values represents networks that have a community structure.

'Edge density' of the trait network (proportion of present connections among traits out of all possible connections) to assess the connectedness across traits. Edge density is used in biological networks, particularly in the study of neural networks, where variation in Edge density has been linked to a compromise between efficiency of connections vs. the cost of connection [125, 7]. The values for Edge density vary between zero to one. A value of zero represents no connection across the traits in the network. Meanwhile a value of one suggests that all traits in the network are connected to all other traits.

Prior to analyses we *log* transformed, calculated species-level means within each climate zone, and z-transformed all continuous trait data. We ran all analyses under R 3.3.1 (R Core Team, 2016). For calculating the precision matrices we used the camel package [124], and for calculating the network metrics we used the igraph package [58].

7.3.3 Results

Among all terrestrial plants, there is a high connection among all traits across all organs (Edge density = 0.86, Figure 7.2). SLA, LLS and SSD were the most central traits, (Degree = 1), followed by seed mass and plant height (Degree = 0.86). Leaf area, leaf N mass and leaf P mass were the least central traits (Degree = 0.71).

Results were generally similar among life forms, but with some significant differences. There were slightly more connections among traits in woody (Edge density = 0.71) than in non-woody species (Edge density = 0.61; Figure 7.2b-c). For non-woody species leaf area, leaf N, LLS and leaf P were the most influential traits in the trait network (Degree = 0.71), while for woody species the most influential traits were leaf area, leaf N and seed mass (Degree = 0.86). In both cases stem related traits were the least influential traits in the network (Degree 0.57; Figure 7.2). Non-woody species showed slightly more connections between traits within organs than between traits among organs (Modularity = 0.10), compared to woody species that show higher integration across organs (Modularity = 0.06). In non-woody species there were two modules, one composed of LLS, leaf N, and SSD, while the other was composed by leaf P, SLA, plant height, leaf area, and seed mass. In woody species we also detected two modules, one

formed by leaf area, leaf P and SSD and the other one by seed mass, LLS, SLA, leaf N and plant height.

In both woody and non-woody species, there were connections among all leaf traits, except SLA-LLS. Also, in both groups of species there were connections between plant height-leaf area, plant height-LLS, seed mass-leaf area, seed mass-SLA, seed mass-leaf N, and seed mass-plant height. Connections between LLS-SLA, SSD-leaf area, SSD-leaf P, plant height-leaf N, seed mass-LLS, and seed mass-SSD were only present among woody species. Connections between SSD-leaf N, SSD-LLS, and plant height-leaf P were only present in non-woody species.

Woody and non-woody species by climate region

Woody species show more connections (i.e. higher Edge density) among traits in tropical (Edge density = 0.64), temperate (Edge density = 0.68), and arid (Edge density = 0.64) environments, compared to cold and polar ones (Edge density = 0.57 and 0.39 respectively; Figure 7.3 and Table 7.9). Also, woody species show slightly higher association of traits within organs in polar (Modularity = 0.31), cold (Modularity = 0.14) and arid environments (Modularity = 0.11), compared to temperate and tropical ones (Modularity = 0.05 and 0, respectively; Figure 7.3; Table 7.9). Non-woody species show more connections between traits in temperate and cold environments (Edge density = 0.64), followed by arid and polar environments (Edge density = 0.43) and least in tropical environments (Edge density = 0.32). Non-woody species show higher modularity in tropical (Modularity = 0.31), arid (Modularity = 0.25) and polar (Modularity = 0.21) environments, while both temperate and cold regions show low modularity (Modularity = 0.07; Table 7.9).

Across the different climate regions woody species always had two modules, while non-woody species had two modules in all climates except polar and tropical ones were they had three (Table 7.3). Both growth forms had a module mainly composed by traits related to the LES. In woody species, this consisted of SLA, leaf N and leaf P. In non-woody species, it contained SLA and leaf N, with leaf P being part of this module in all climate regions except tropical and temperate climates. The second module for both growth forms was composed of traits related to the reproductive strategy and plant architecture. In woody species, the core traits in this module were seed mass, plant

height and leaf area, with SSD being part of this module in all climate regions except temperate areas. In non-woody species this module was composed of plant height and leaf area, SSD was part of this module in all climates except polar. When a third module was present in non-woody species, LLS and either leaf P in tropical, or SSD in polar climates composed it (Table 7.3).

In terms of centrality of traits, as measured by their Degree, for both growth forms LLS was a central trait in temperate, arid and cold climates, while seed mass was a central trait in tropical areas (Table 7.8). For non-woody species leaf area was central in all climate zones, except temperate and polar, while leaf N was central in all regions except temperate and cold ones (Table 7.8).

Across climate types and both growth forms, connections between leaf N-leaf P, leaf N-SLA, plant height-leaf area and seed mass-leaf area were always present (Figure 7.3). For non-woody species connections among seed mass-SLA, and seed mass-leaf N were also robust across climate types (Figure 7.3f-j). Meanwhile, for woody species connections between SLA-leaf area, leaf P-SLA, LLS-leaf N, SSD-leaf P, seed mass-plant height, and seed mass-SSD were also found across climates regions (Figure 7.3a-e).

Analyses using area-based Leaf N and P metrics produced results largely consistent with those using mass based leaf nutrient content measurements in terms of trait connections and modularity (Table 7.9; Figures 7.4, 7.5). The robustness of connections across climates and growth forms was also similar between mass and area based results, with the addition of a connection between LLS-SLA in the area based results. As in the mass based results, no unique trait was central across all climates for both woody and non-woody area-based traits (Table 7.8). However, the Degree of traits did change within climates (Table 7.8). For instance, in woody species LLS became a central trait in all climates except polar. For both groups seed mass stopped being a central trait in tropical areas. Similarly, leaf area was no longer a central trait in non-woody species. Meanwhile, SSD became a central trait for woody species from arid regions (Table 7.8).

7.4 Discussion

In the current study, we used precision matrices of a large global dataset of ten traits that represent all above-ground plant organs (leaf, stem and reproductive) across 16,281

plant species to identify the direct connections that exist across traits across organs for this ten-trait constellation. We identified emergent characteristics of the trait networks across all land plants at a global scale, as well as across growth forms (woody and non-woody species) and, for the first time (to our knowledge), explicitly accounted for the impact that broad environmental gradients have on the trait network. In doing so, our study builds on and extends previous attempts that describe the cross-correlations across several traits and several organs at global scales [64], and efforts that focused on certain vegetation types (e.g. [223]) and narrower environmental gradients (e.g. [50]). Important steps forward in this type of analyses will be the incorporation of: belowground traits once global root trait databases have grown to a significant cover to investigate the generality of the connection among LES- root economic spectrum traits across regions (e.g. [127, 74, 50, 121]), and that of whole-plant leaf area data to investigate the linkage between lower SLA, longer LLS and whole-plant leaf area [153].

7.4.1 Connectivity across all terrestrial plants

At a global scale we found that land plants have high connectivity across traits and high integration across organs (Figure 7.2a). This supports the idea that matching tissue strategies should be advantageous at the whole-plant scale [168]. In terms of traits, SLA, LLS, and SSD were the most central traits across land plants (Figure 7.2a). High centrality suggests that a variable tends to be influential in terms of regulating critical functions or being involved in the regulation of more functions, and therefore of having greater impact on higher level properties, such as fitness [107]. Indeed, SLA and LLS are crucial traits in the resource acquisition strategy of plants, representing the compromise between the carbon construction cost and the duration of this benefit [167, 171, 170], while SSD impacts plant hydraulic and mechanical properties and influences the nutrient, carbon and water economy of stems [45]. Further, variation in these three traits has been shown to impact growth and fecundity, while also contributing to the structuring of communities [111] and heavily influencing ecosystem level processes [171, 168].

7.4.2 Trait connections across growth forms and climate regions

Connections between seed mass-leaf area, leaf N-SLA, leaf P-leaf N, and plant height-leaf area were always present in analyses of growth forms, and of growth forms across different climate types (Table 7.4). Some of these connections are well known (e.g. leaf N-SLA, leaf P-leaf N), and previous correlation analyses have identified their importance in understanding compromises among traits and their impact on plant function [167, 171, 170, 224]. Some others are connections that previous correlation analyses suggest have no overall or weak relationships across habitats, and therefore their importance has been downplayed (e.g. seed-mass-leaf area, plant height-leaf area; Table 7.4). Thus, correlation strength does not reflect connectivity between traits well (Tables 7.5, 7.6, 7.7). In the first case above where we detect a connection and previous studies show a strong correlation evidence suggests that these connections are maintained through selective pressure of biophysical constraints and natural selection (Table 7.4). For example, an increase in SLA will generally be linked to an increase in leaf N and other cytoplasmic molecules [171, 143]. At the same time, natural selection reinforces a strong connection between SLA-leaf N relationship through processes such as herbivory and competition [171], limiting the trait space where optimal combinations of these traits occur. In the second case where conditional dependency between two traits exists, but previous correlation tests suggest a weak relationship- we propose that these trait connections are maintained in the plant phylogeny through neutral or selective processes, but contradictory selective forces across habitats weaken the correlation among these traits. For instance, studies have reported a triangular relationship between seed mass and leaf area in temperate woody species (i.e. big leaves have big or small seeds, but small leaves only have small seeds; [55]). But positive rather than triangular relationships across woody sclerophyll species [220], and no relationship among these traits across woody tropical species have been reported as well [223]. In both cases, our analyses suggest that these trait connections are robust after accounting for all other traits, and across climates and growth forms, but the forces maintaining these connections might differ.

Some well-known, strong trait-trait correlations across plants have robust connections globally but not across growth forms or climate regions in our study. For instance, the connection between SLA-LLS (which additionally are two of the three most central

traits globally) is direct in the global data and in woody species (globally) but not for non-woody plants (globally). Moreover, this connection was observed in four of five climate regions for woody plants, but only one of five for non-woodies. This weaker connection in subsets of the global data could suggest that although a strong correlation exists between these two traits even in these subsets, the connection between these traits could be mechanistically mediated through other traits in some cases, or the connection only exists when the absolute range in LLS is large, as is the case for woody plants.

7.4.3 Modularity

In biological or man-made networks, modules are a group of nodes that interact more strongly among themselves, and tend to perform a common function [7]. We found that in woody species across climate regions, SLA, leaf N, and leaf P were always together in a module. These three physiological traits are central to the leaf economic spectrum [224]. Seed mass (a reproductive trait) and plant height and leaf area (architectural traits) formed the core traits in a second module (Table 7.3). These traits are relevant to body size, plant-water, and -light relations [154, 217, 147]. Similarly, in non-woody species, leaf N and SLA were always together in one module (leaf P was part of this module, except in tropical climates), while plant height and leaf area were together in a separate module (SSD formed part of this module except in polar climate). Our results suggest that there are at least two core modules across plant growth forms, one module whose function is carbon uptake represented by leaf physiological traits, and a second module whose function is more related to body size, and plant-light, -water relations represented by seed mass and two architectural traits. The importance of these two modules in the strategies of plants is supported by the LES and the global spectrum of plant form and function [171, 170, 64].

7.4.4 Modularity across climate regions and growth forms

On average we found higher modularity in non-woody species than in woody species. Modularity evolves in more variable environments [125] because it provides robustness

against component failures [7]. This result may provide some insight into the observation that non-woody species tend to succeed in drier, colder, and more stochastic environments, compared to woody species [113, 148], and provides some support to the observation that shrubs have more segmented hydraulic stem design under drier conditions [177]. Within woody species modularity increases from tropical to polar climates, consistent with the idea of modularity as providing robustness in stressful or variable environments. Interestingly, non-woody species showed higher modularity in tropical, polar and arid regions (Figure 7.3h-j). Strong modularity of non-woody tropical could be explained by the variable precipitation that seasonal tropical climates experience, while high modularity in arid and polar climates could be related to seasonal patterns in temperature, precipitation or both. The seemingly counterintuitive pattern in modularity of non-woody species may be partially explained by difference in the geographic sampling bias of the two growth forms and the composition of tropical climate in our study. In this study tropical climate includes not only rainforest, but also tropical seasonal forest and savannahs, which experience high rainfall seasonality. The majority of data on tropical non-woody species comes from savannah and tropical seasonal forest, while a greater proportion of the data for tropical woody species comes from rainforest.

7.4.5 Modularity across a precipitation and a temperature gradient

Higher variation in modularity in non-woody species across climate regions could also be partially explained by geographic bias in the distribution of monocots and eudicots species across climate regions and differences in morphology, development, tissue longevity, and anatomy of these clades [56, 203]. Thus, we ran some exploratory analyses to assess the effect that temperature (holding precipitation constant) and precipitation (holding temperature constant) have on the modularity of non-woody forbs and monocots, as well as across woody angiosperms. These exploratory analyses revealed that in drier conditions woody angiosperms show an increase in modularity, while forbs show an increase in modularity in wetter environments (Tables 7.11, 7.10; Figures 7.7, 7.9). Monocots show higher modularity under dry and wet conditions compared to intermediate levels of precipitation. Forbs also show higher modularity in colder climates. Taken together, these results suggest that fewer connections between tissues might be more advantageous in environments that experience more variable conditions, which may lead

to greater temporal asynchrony between resources [24], and therefore higher modularity in the plant strategies.

In summary, this approach and these analyses help increase our understanding of the mechanisms behind trait relations by studying their statistical dependencies in uncollapsed trait space. Probabilistic graphical models describe sets of causal hypotheses that can be used to further investigate cause-effect relationships through structural equation models (SEM; [135]. Thus, the dependencies among traits provided by the probabilistic graphical model developed in our study can be further explored with SEMSs to understand the coordination, compromises and causation across multiple traits (e.g. as for leaf traits alone [185]). Further, using the probabilistic graphical models represented by the precision matrix one may be able to estimate the number of latent variables missing. This is possible by leveraging the sparsity of the precision matrix in combination with a low-rank matrix, where the rank of the low-rank matrix represents the number of missing latent variables ([41]; Section 7.4.7). These issues will not be covered here, but should be considered as areas of research with potential to advance our understanding of the trait relationships that govern plant strategies.

Previous plant trait studies have focused on understanding the interdependence between pairs of traits or among multiple traits using correlation analyses in combination with ordination techniques (e.g. Table 7.1). Here we present an analysis of the conditional dependencies among multiple traits across land plants, growth forms and climate regions that allow us to differentiate direct from indirect interactions among traits. The findings presented here contribute to the fundamental understanding of dependencies between plant traits across environmental gradients. We found that SLA, LLS and SSD are the most central traits globally, but were not always connected within climate zones, especially for non-woody plants; moreover few trait-trait connections exist robustly across all growth forms and climate gradients (Table 7.4). Surprisingly, we found that not all strong correlations are direct connections, and some weak correlations are direct connections (Tables 7.5, 7.6, 7.7). Despite the difference in statistical approach with previous research, our study supports the existence of two distinguishable dimensions or functional modules across land plants and climate regions. One module is related to physiological leaf traits related to carbon uptake and economy, and another

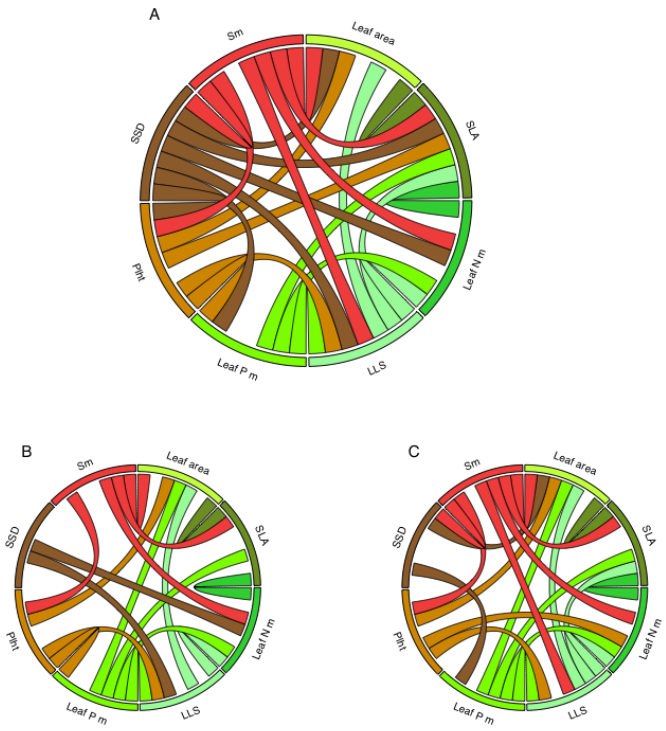


Figure 7.2: Connections between multiple traits across organs (leaves in greens, stems in browns and seed in red) using mass units for leaf N and P content. (A) all terrestrial plants, (B) non-woody species and (C) woody species.

related to reproductive strategy and plant architecture. The approach taken here represents an important step forward on the collective path to understand the causal links among multiple traits across multiple organs and within and across different climate zones and plant life forms.

7.4.6 Plant trait network analyses on a precipitation and temperature gradient holding the other environmental variable constant across each gradient

We used the CRU TS 0.5 gridded climate dataset (Harris et al., 2014) to derived mean annual temperatures, and mean annual precipitation data for our georeferenced plant records. Then, we created four precipitation bins (0-500 mm, 500-1200mm, 1200-2500

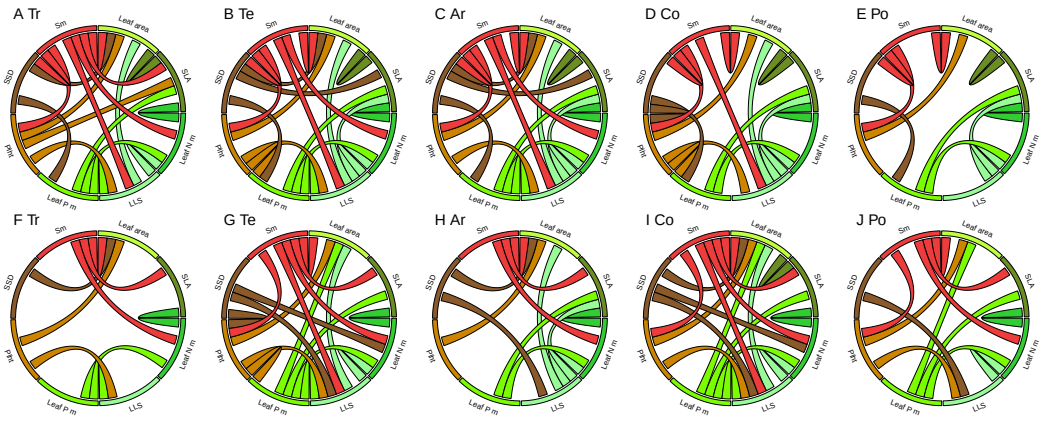


Figure 7.3: Connections between traits across organs (leaves in greens, stems in browns and seed in red) for woody (A-E) and non-woody species (F-J) in (Tr) Tropical, (Te) Temperate, (Ar) Arid, (Co) Cold, and (Po) Polar environments. Environment types were derived using the Kppen Climate Zones classification system (Peel et al., 2007; see methods).

mm, 2500 mm, holding temperature between 10-20 C), and four temperature bins (0 C, 0-10 C, 10-20 C, 20 C, holding precipitation between 0-500 mm). For the precipitation bins, data availability made it only possible to run the analyses on three bins (0-500 mm, 500-1200mm, 1200-2500mm). We divided the plant records into woody angiosperms, woody gymnosperms, non-woody forbs and non-woody monocots. Then we obtained the precision matrices for these groups of species across our precipitation and temperature bins define above. We excluded non-woody monocots from the temperature analyses and woody gymnosperms from both the precipitation and temperature analyses as we lacked enough data for these groups across the environmental gradients. Results for these comparisons are shown below.

7.4.7 Calculation of latent variables using a sparse precision matrix and low rank matrix

Current model assumes that we have fully observed all variables. However, this assumption may not be a valid assumption. For instance in our case there might be other

important traits that are being neglected in the current analysis. For example, in the example below, wet street and wet grass are connected in the conditionally dependency graph when the variable rain is not observed. But when the variable rain is observed, wet street and wet grass are are disjoint (Figure 7.10).

In another example, in the presence of a sprinkler the interaction structure may change to the graph in Figure 7.10c. Here, wet street and wet grass are conditionally independent given both rain and sprinkler as rain and sprinkler can explain all relationship that may exist between wet street and wet grass. Thus, finding the graph structure between observed traits while considering presence of latent variables is of interest and important.

In general, consider we have measured p traits $X = (X_1, X_2, \cdots, X_p)$ of a $p + r$ multivariate Gaussian vectors of traits (X, Y) where $Y = (Y_1, Y_2, \cdots, Y_r)$ represents the latent traits. It is natural to assume that the fully observed $p + r$ traits have a sparse precision matrix. But, since we have not observed r of traits, we cannot apply the current model to obtain such sparse network. Interestingly, it has been shown that the interaction networks between p observed traits can be represented as sum of a sparse A and a low rank matrix B where matrix B has a rank at most r which is number of latent traits [41].

Table 7.1: Examples of studies focused on multi-organ, multi-trait datasets. When several plant group classifications were used, a semicolon divides them. The numbers next to the name of the organs included the study in the Organs column refers to the number of traits per organ. For more details on these studies see Appendix S.

study	**Growing conditions of the plants**	**Spatial scale**	**Environmental gradients**	**Are environmental gradients explicitly consider?**	**Plant groups included**	**Plant groups taken into consideration?**	**Number of species**	**Life stage**	**Organs**	**Observation**
Ackerly 2004	field	local	Water availability	no	woody; evergreen, woody deciduous	yes	20	adults	leaf 23 stem 11 seed 2	Effect of water availability inferred from plant water potentials
Baraloto et al. 2010	field	regional	Precipitation, Soil nutrients	no	trees	no	668	adults	leaf 12 stem 5	
Cheng et al. 2015	field	local	Temperature, Precipitation, Soil nutrients	yes	xerophytes, intermediate xerophytes, intermediate mesophytes, mesophytes shallow vs deep-rooted	yes	55	adults	leaf 1 roots 4	
de la Riva et al. 2015	field	local	Humidity (soil water)	yes	woody		38	adults	leaf 7 stem 1 roots 5 seed 1	
Diaz et al. 2004	field	regional	Precipitation, Temperature, Altitude	no	eudicotyledons, monocotyledons; asteraceae; fabaceae; poaceae	yes	640	adults	leaf 5 stem 4 seed 2 other 2	
Diaz et al. 2016	field	global	Precipitation, Temperature, Altitude	no	woody, non-woody; angiosperms, gymnosperms, pteridophyte	yes	46085	adults		
Fortunel et al. 2012	field	regional	Precipitation, Soil nutrients	yes	trees	no	758	adults	leaf 11 stem 2 root 1	
Freschet et al. 2010	field	local	soil nutrients (C &N), Temperature (soil litter), Humidity (soil litter), growing season	yes	woody evergreen, woody deciduos, fern allies, club mosses, monocots, terretrial forbs, aquatic forbs	yes	40	adults	leaf 8 stem 7 roots 7	
Ishida 2008	field			no	trees, creeping trees, ruderal trees, climber, C3 shrubs and forbs, CAM, palm		32	adults	leaf 14 stem 1	
Jager et al. 2015	field	local	Terrain slope, Soil nutrients	yes	angiosperm trees, conifer trees, palm trees, fern trees	yes	30	> 10 dbh	leaf 8 stem 4 seed 1	
Kramer et al. 2016	glasshouse	local, regional	Soil nutrients	yes	Conifer, Eudicot, Palm, Magnoliid, Tree fern; AM(nodules), AM, Non, Dual AM/EM, EM, Ericoid	yes	66	seedlings	leaf 2 stem 2 root 5	Plants sourced from nurseries; traits from seedling from glasshouse
Li and Bao 2015	common garden				woody deciduos, woody xerophythic	no	23	1 year old	leaf 3 stem 1 roots 4	
Wright 2007	field	regional	Precipitation	no	trees, shrubs, lianas	no	122	adults	leaf 2 stem 2 seed 2	main analyses done with 122 species, supplementqry analyses done with up to 2134 spp

Table 7.2: Comparison of trait-trait correlations using only gap-filled or only original trait values. Sample sizes (n) of the gap filled and original database are the same to ensure they are comparable. To test whether the gap-filling algorithm had an effect on trait-trait correlation we used a subset of 470769 observations from TRY for five traits (leaf area, SLA, leaf n, plant height and seed mass). First we ran the gap filling algorithm on this dataset. Then using standardize major axis analyses we compared the trait-trait correlations of a dataset only using observed values vs. the exact same observations only using gap filled values. The sample sizes for these trait-trait correlations varied between 1738 for the leaf N-seed mass correlation to 63846 records for the SLA-leaf area correlation. Overall, the difference in the slope value of the trait-trait correlation between the two types of data (gap-filled vs original) ranged from 0.0005 to 0.06, and in the case of the intercepts it varied between 0.006 and 0.11.

Correlation	Data type	n	Slope	Slope low CI	Slope high CI	Intercept	Intercept low CI	Intercept high CI
Leaf area-height	Gap filled	28278	1.0057	0.9962	1.0153	6.8396	6.8189	6.8603
	Original	28278	1.0059	0.9961	1.0157	6.8461	6.8241	6.8681
Leaf area-leaf n	Gap filled	9154	4.6452	4.552	4.7402	-6.0163	-6.2988	-5.7337
	Original	9154	4.5885	4.4964	4.6824	-5.8969	-6.1777	-5.6161
Leaf area-seed mass	Gap filled	1693	0.7785	0.7445	0.8141	5.9082	5.7919	6.0245
	Original	1693	0.81	0.7742	0.8475	5.9834	5.8582	6.1085
Leaf n-plant height	Gap filled	9013	-0.1616	-0.165	-0.1583	3.2256	3.2155	3.2356
	Original	9013	-0.1716	-0.1752	-0.1681	3.2524	3.2415	3.2632
Leaf n-seed mass	Gap filled	1738	0.1344	0.1283	0.1409	3.0283	3.0069	3.0498
	Original	1738	0.1418	0.1353	0.1486	3.0365	3.0135	3.0596
Plant height-seed mass	Gap filled	5129	0.4494	0.6205	0.6463	-0.6422	-0.6794	-0.6049
	Original	5129	0.3997	0.6248	0.6519	-0.5556	-0.5957	-0.5154
SLA-leaf area	Gap filled	63846	0.322	0.3195	0.3244	0.5228	0.5045	0.5411
	Original	63846	0.3297	0.3272	0.3322	0.475	0.4562	0.4938
SLA-leaf n	Gap filled	20044	0.2938	1.446	1.48	-1.785	-1.8357	-1.7344
	Original	20044	0.2949	1.4229	1.4563	-1.7541	-1.8042	-1.704
SLA-plant height	Gap filled	33212	-0.3448	-0.3484	-0.3413	2.8135	2.8048	2.8222
	Original	33212	-0.3535	-0.3572	-0.3499	2.7928	2.7836	2.8021
SLA-seed mass	Gap filled	3256	-0.2458	-0.2543	-0.2376	3.0297	3.0041	3.0553
	Original	3256	-0.2558	-0.2646	-0.2472	3.022	2.9942	3.0497

Table 7.3: Modules of non-woody and woody species across climate regions. The pipe character '|' separates individual modules. Traits across modules may be connected (see Figure 7.3), however they tend to be more connected with other traits within the modules than with traits outside the module.

Climate	Woody species modules	Non-woody species modules
tropical	SLA-leaf N-leaf P | plant ht-leaf area-seed mass-ssd-lls	SLA-leaf N-seed mass | plant ht-leaf area-SSD | LLS-leaf p
temperate	SLA-leaf N-leaf P-ssd | plant ht-leaf area-seed mass-lls	SLA-leaf N-seed mass | plant ht-leaf area-seed mass-lls-leaf p |
arid	SLA-leaf N-leaf P-LLS | plant ht-leaf area-seed mass-ssd	SLA-leaf N-leaf P-seed mass | plant ht-leaf area-SSD-LLS |
cold	SLA-leaf N-leaf P | plant ht-leaf area-seed mass-ssd-lls	SLA-leaf N-leaf P | plant ht-leaf area-seed mass-ssd-lls |
polar	SLA-leaf N-leaf P-LLS | plant ht-leaf area-seed mass-ssd	SLA-leaf N-leaf P | plant ht-leaf area-seed mass | LLS-SSD

Table 7.4: Trait connections that are robust (i.e. common across groups) across growth forms and climate regions and proposed mechanisms that maintain this connection. The range of R^2 values observed across growth form, and then by growth form across climate regions in this study is provided from the second to fourth columns. We provided mechanism proposed to maintain these trait connection as well as specific hypothesis about this mechanism (Details column).

Robust connection	**Woody/ non-woody**	**Woody across climate**	**Non-woody across climate**	**Mechanisms**	**Details**
leaf P-leaf N	0.62	0.51-0.70	0.47-0.72	biochemical and biogeochemical constraints. Selection through competition and herbivory	Selection through biochemical constraints reinforced through compromises between allocation of resources to metabolism (investment in genetic material esp. ribosomal RNA) vs growth (investment in proteins); biogeochemical constraints mediated by adaptation/acclimation to soil conditions; unviable/low fitness strategies may be selected out by herbivory and competition based processes (Reich & Oleksyn, 2004; Kerkhoff et al., 2006; Reich, 2014)
SLA-leaf N	0.52-0.64	0.59-0.67	0.41-0.57	biophysical constraint; selection through competition, herbivory of viable strategies.	Changes in leaf N results from increases in LMA, increase in SLA result in an increase in cytoplasmic compound, including N; structural investment, competition and herbivory select certain combinations of these traits (Reich et al., 1992; Reich et al., 1997; Reich et al., 1999; Meziane & Shipley, 2001)
plant height-leaf area	0.50-0.54	0.29-0.57	0.36-0.70	biophysical constraint	Decrease in leaf size with increasing light demands correlate weakly, selection through light competition in certain florae; big leaf -small trunk physically unviable; large-leaf crown is more efficient because it requires less woody support investment (Givnish, 1979; Niinemets & Kull, 1994; Niklas, 1994)
Seed mass-leaf area	0.41-0.42	0.15-0.50	0.36-0.41	unclear if this correlation is general across floras or what maintains it. Possible controls by biophysical constraints or vascular/meristematic demands	Correspondence between axes size and appendages size (Corner's rule). Triangular relationship, plants with big leaves can have either very small or large seeds, but plants with small leaves only have small seeds in woody European species; positive no triangular relationship in sclerophyll vegetation in Australia; no relationship among these traits in tropical forest in America (Cornelissen, 1999; Westoby & Wright, 2003; Wright et al., 2007)

References for table: Cornelissen JHC. 1999. A triangular relationship between leaf size and seed size among woody species: allometry, ontogeny, ecology and taxonomy. Oecologia 118(2): 248-255; Givnish T 1979. On the adaptive significance of leaf form. Topics in plant population biology: Springer, 375-407; Kerkhoff AJ, Fagan WF, Elser JJ, Enquist BJ. 2006. Phylogenetic and growth form variation in the scaling of nitrogen and phosphorus in the seed plants. The American naturalist 168(4): E103-E122; Meziane D, Shipley B. 2001. Direct and indirect relationships between specific leaf area, leaf nitrogen and leaf gas exchange. Effects of irradiance and nutrient supply. Annals of Botany 88(5): 915-927; Niinemets Ü, Kull K. 1994. Leaf weight per area and leaf size of 85 Estonian woody species in relation to shade tolerance and light availability. Forest Ecology and Management 70(1-3): 1-10; Niklas KJ. 1994. Plant allometry: the scaling of form and process: University of Chicago Press; Reich PB. 2014. The worldwide fastslowplant economics spectrum: a traits manifesto. Journal of Ecology 102(2): 275-301; Reich PB, Ellsworth DS, Walters MB, Vose JM, Gresham C, Volin JC, Bowman WD. 1999. Generality of leaf trait relationships: a test across six biomes. Ecology 80(6): 1955-1969; Reich PB, Oleksyn J. 2004. Global patterns of plant leaf N and P in relation to temperature and latitude. Proceedings of the National Academy of Sciences of the United States of America 101(30): 11001-11006; Reich PB, Walters MB, Ellsworth DS. 1992. Leaf lifespan in relation to leaf, plant, and stand characteristics among diverse ecosystems. Ecological Monographs 62(3): 365-392; Reich PB, Walters MB, Ellsworth DS. 1997. From tropics to tundra: global convergence in plant functioning. Proceedings of the National Academy of Sciences 94(25): 13730-13734; Westoby M, Wright IJ. 2003. The leaf sizetwig size spectrum and its relationship to other important spectra of variation among species. Oecologia 135(4): 621-628; Wright IJ, Ackerly DD, Bongers F, Harms KE, Ibarra-Manriquez G, Martinez-Ramos M, Mazer SJ, Muller-Landau HC, Paz H, Pitman NCA. 2007. Relationships among ecologically important dimensions of plant trait variation in seven Neotropical forests. Annals of Botany 99(5): 1003-1015.

Table 7.5: Trait-trait correlations (r) and precision matrix values (ω) for all land plants, woody and non-woody species.

	All plants		**Woody**		**Non-woody**	
Trait-trait connection	**r**	**ω**	**r**	**ω**	**r**	**ω**
Leaf area-SLA	0.07	-0.33	0.37	-0.54	0.07	-0.03
Leaf area-LeafN	0.11	0	0.25	0	0.13	0
Leaf area-LLS	0.11	0.25	-0.1	0.19	-0.12	0.17
Leaf area-LeafP	-0.05	0	0.11	0.19	0.1	-0.05
Leaf area-PtH	0.6	-1.06	0.54	-0.75	0.51	-0.71
Leaf area-SSD	0.11	0.14	-0.16	0.24	-0.04	0
Leaf area-Sm	0.52	-0.5	0.42	-0.4	0.41	-0.34
SLA-Leaf N	0.59	-0.65	0.64	-0.69	0.52	-0.54
SLA-LLS	-0.54	0.2	-0.53	0.32	-0.29	0
SLA-Leaf P	0.62	-0.6	0.6	-0.58	0.5	-0.4
SLA-Plant height	-0.27	0.04	0.06	0	-0.08	0
SLA-SSD	-0.32	0.09	-0.28	0	0.02	0
SLA-Seed mass	-0.26	0.36	0	0.01	-0.07	0.25
Leaf N-LLS	-0.56	0.73	-0.57	0.69	-0.48	0.47
Leaf N-Leaf P	0.61	-0.69	0.62	-0.65	0.62	-0.76
Leaf N-Plant height	-0.14	0	0.07	-0.05	-0.14	0
Leaf N-SSD	-0.16	-0.06	-0.2	0	0.05	-0.05
Leaf N-Seed mass	0	-0.51	0.08	-0.27	0.23	-0.32
LLS-Leaf P	-0.58	0.34	-0.54	0.33	-0.4	0.15
LLS-Plant height	0.5	-0.66	0.25	-0.36	0.23	-0.28
LLS-SSD	0.38	-0.13	0.21	0	0.1	-0.08
LLS-Seed mass	0.41	-0.38	0.28	-0.41	-0.13	0
Leaf P-Plant height	-0.33	0.08	-0.06	0	-0.18	0.12
Leaf P-SSD	-0.35	0.21	-0.36	0.35	-0.01	0
Leaf P-Sm	-0.25	0	-0.11	0	0.12	0
Plant height-SSD	0.37	-0.2	0.03	0	0	0
Plant height-Seed mass	0.68	-0.84	0.55	-0.55	0.31	-0.22
SSD-Seed mass	0.39	-0.31	0.19	-0.26	0.01	0

Table 7.6: Woody species trait-trait correlations (r) and precision matrix values (ω) across climate regions

	Tropical		**Temperate**		**Arid**		**Cold**		**Polar**	
Trait-trait connection	**r**	**ω**	**r**	**ω**	**r**	**ω**	**r**	**ω**	**r**	**ω**
Leaf area-SLA	0.2	-0.21	0.36	-0.26	0.36	-0.39	0.58	-0.71	0.52	-0.34
Leaf area-LeafN	0.13	0	0.28	0	0.19	0	0.38	0	0.35	0
Leaf area-LLS	-0.05	0.002	-0.26	0.3	-0.16	0.18	-0.46	0.47	-0.42	0
Leaf area-LeafP	0.1	0	0.24	0	0.1	0	0.34	0	0.31	0
Leaf area-PtH	0.29	-0.29	0.43	-0.5	0.57	-0.77	0.3	-0.26	0.37	-0.11
Leaf area-SSD	-0.22	0.2	-0.2	0.19	-0.17	0.1	0.04	0	0.07	0
Leaf area-Sm	0.16	-0.1	0.39	-0.4	0.5	-0.51	0.37	-0.33	0.36	-0.1
SLA-Leaf N	0.6	-0.73	0.67	-0.5	0.64	-0.48	0.59	-0.38	0.64	-0.32
SLA-LLS	-0.43	0	-0.64	0.61	-0.64	0.67	-0.63	0.51	-0.64	0.51
SLA-Leaf P	0.55	-0.44	0.68	-0.74	0.61	-0.35	0.52	-0.38	0.57	-0.28
SLA-Plant height	-0.15	0.05	0.02	0	0.03	0	-0.05	0	-0.05	0
SLA-SSD	-0.26	0	-0.32	0.08	-0.38	0.18	0	0	-0.09	0
SLA-Seed mass	-0.2	0.12	0.02	0	0.01	0	0.06	0	-0.04	0
Leaf N-LLS	-0.51	0.59	-0.66	0.83	-0.63	0.59	-0.63	0.73	-0.66	0.6
Leaf N-Leaf P	0.55	-0.44	0.7	-0.94	0.7	-0.97	0.51	-0.35	0.65	-0.64
Leaf N-Plant height	-0.05	0	0.05	0	0.02	0	0.03	0	0.01	0
Leaf N-SSD	-0.16	0	-0.25	0	-0.23	0	-0.03	0	-0.2	0
Leaf N-Seed mass	-0.02	-0.24	0.09	-0.15	0.08	-0.1	0.06	0	-0.06	0
LLS-Leaf P	-0.46	0.26	-0.58	0.18	-0.6	0.31	-0.43	0	-0.46	0
LLS-Plant height	0.28	-0.17	0.17	-0.32	0.23	-0.34	0.22	-0.28	0.18	0
LLS-SSD	0.2	0	0.22	0	0.32	0	-0.04	0	0.05	0
LLS-Seed mass	0.38	-0.43	0.14	-0.21	0.21	-0.22	0.1	-0.08	0.08	0
Leaf P-Plant height	-0.13	0	0.07	-0.05	-0.07	0	0.12	-0.02	0.11	0
Leaf P-SSD	-0.36	0.34	-0.35	0.26	-0.39	0.26	-0.17	0.1	-0.3	0.06
Leaf P-Sm	-0.18	0	0.02	0	-0.06	0	0.07	0	-0.08	0
Plant height-SSD	0.06	0	0.01	0	0.08	0	-0.07	0.11	-0.02	0
Plant height-Seed mass	0.43	-0.41	0.47	-0.42	0.55	-0.44	0.47	-0.48	0.43	-0.21
SSD-Seed mass	0.21	-0.18	0.2	-0.27	0.16	-0.14	0.33	-0.34	0.42	-0.23

Table 7.7: Non-woody species trait-trait correlations (r) and precision matrix values (ω) across climate regions

	Tropical		**Temperate**		**Arid**		**Cold**		**Polar**	
Trait-trait connection	**r**	**ω**	**r**	**ω**	**r**	**ω**	**r**	**ω**	**r**	**ω**
Leaf area-SLA	0.1	0	-0.003	0	0.12	0	0.16	-0.03	0.08	0
Leaf area-LeafN	0.2	0	0.15	0	0.13	0	0.19	0	0.25	0
Leaf area-LLS	-0.02	0	-0.19	0.19	-0.21	0.05	-0.23	0.13	-0.11	0
Leaf area-LeafP	0.13	0	0.12	0	0.09	0	0.24	-0.11	0.25	-0.06
Leaf area-PtH	0.36	-0.29	0.45	-0.57	0.36	-0.31	0.55	-0.74	0.6	-0.68
Leaf area-SSD	0.2	-0.06	-0.06	0	-0.22	0.11	-0.12	0.06	-0.07	0
Leaf area-Sm	0.42	-0.36	0.39	-0.27	0.36	-0.29	0.39	-0.27	0.5	-0.32
SLA-Leaf N	0.56	-0.62	0.57	-0.68	0.49	-0.35	0.43	-0.37	0.41	-0.24
SLA-LLS	-0.28	0	-0.31	0	-0.38	0.16	-0.18	0	-0.21	0
SLA-Leaf P	0.46	0	0.54	-0.42	0.48	-0.32	0.42	-0.31	0.42	-0.27
SLA-Plant height	0.08	0	-0.14	0	0.002	0	0.04	0	0.07	0
SLA-SSD	-0.04	0	0.02	0	-0.08	0	0.08	0	0.04	0
SLA-Seed mass	0.02	0.03	-0.12	0.25	-0.08	0.05	0	0.11	-0.05	0.02
Leaf N-LLS	-0.33	0	-0.51	0.5	-0.46	0.36	-0.41	0.36	-0.44	0.4
Leaf N-Leaf P	0.72	-1.25	0.64	-0.77	0.59	-0.67	0.48	-0.41	0.47	-0.36
Leaf N-Plant height	0.03	0	-0.12	0	-0.07	0	0.03	0	0.24	0
Leaf N-SSD	0	0	0.07	-0.02	0.02	0	0.1	-0.08	-0.05	0
Leaf N-Seed mass	0.36	-0.19	0.2	-0.28	0.14	-0.06	0.31	-0.3	0.36	-0.23
LLS-Leaf P	-0.35	0.13	-0.42	0.15	-0.36	0	-0.3	0.09	-0.23	0
LLS-Plant height	0.26	-0.17	0.2	-0.25	0.06	0	0.02	-0.07	0.09	-0.05
LLS-SSD	-0.06	0	0.07	-0.02	0.23	-0.12	0.12	-0.08	0.23	-0.09
LLS-Seed mass	-0.05	0	-0.18	0.07	-0.02	0	-0.24	0.07	-0.09	0
Leaf P-Plant height	-0.09	0	-0.15	0.01	-0.09	0	0.03	0	0.14	0
Leaf P-SSD	-0.08	0	0.02	0	-0.03	0	0.04	0	-0.12	0
Leaf P-Sm	0.32	0	0.11	0	-0.03	0	0.16	0	0.12	0
Plant height-SSD	0.02	0	-0.07	0.01	-0.06	0	-0.06	0	0	0
Plant height-Seed mass	0.14	0	0.3	-0.22	0.23	0	0.31	-0.14	0.5	-0.34
SSD-Seed mass	0.15	0	0.01	0	-0.03	0	-0.02	0	-0.13	0

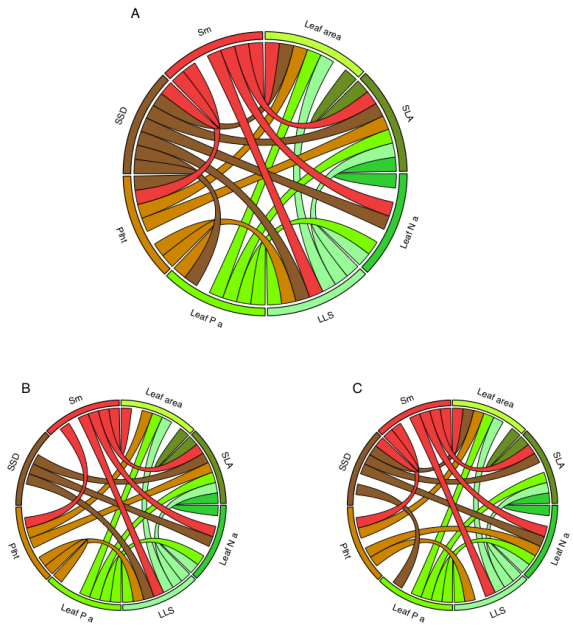


Figure 7.4: Connections between traits across organs (leaves, stems and seed) using area units for leaf N and P content. (A) All terrestraial plants included in this study, (B) No-woody species, (C) Woody species.

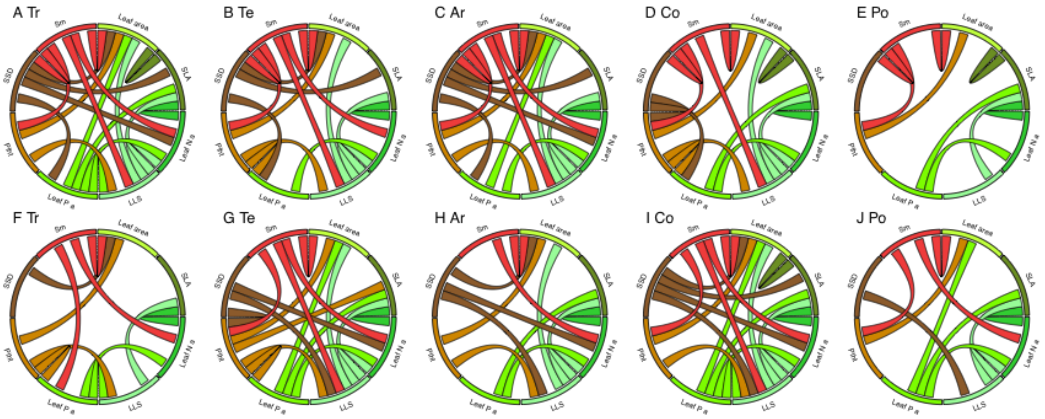


Figure 7.5: Connections between traits across organs (leaves, stems and seed) for woody (A-E) and non-woody species (F-J) across a climate gradient and using area-based measurements of leaf N and P content. The climate regions are Tr Tropical, Te Temperate, Ar Arid, Co Cold, Po Polar.

Table 7.8: Trait centrality (i.e. degree) for woody and non-woody species grouped into five different climate regions. Analyses were ran using (A) mass-based, and (N) area based leaf nutrient content (N, and P).

	Climate	Group	Leaf area	SLA	Leaf N	LLS	Leaf P	Plant height	SSD	Seed mass
(A) Mass based results	Tropical	Woody	0.71	0.71	0.57	0.71	0.57	0.57	0.43	**0.86**
	Temperate	Woody	0.71	0.71	0.57	**0.86**	0.71	0.57	0.57	0.71
	Arid	Woody	0.71	0.71	0.57	**0.86**	0.57	0.43	0.57	0.71
	Cold	Woody	0.57	0.57	0.43	**0.71**	0.57	**0.71**	0.43	0.57
	Polar	Woody	0.43	**0.57**	0.43	0.29	0.43	0.29	0.29	0.43
	Tropical	Non-woody	**0.43**	0.29	**0.43**	0.29	0.29	0.29	0.14	**0.43**
	Temperate	Non-woody	0.57	0.43	0.71	**0.86**	0.71	0.71	0.43	0.71
	Arid	Non-woody	**0.57**	**0.57**	**0.57**	**0.57**	0.29	0.14	0.29	0.43
	Cold	Non-woody	**0.86**	0.57	0.71	**0.86**	0.57	0.43	0.43	0.71
	Polar	Non-woody	0.43	0.43	**0.57**	0.43	0.43	0.43	0.14	**0.57**
(B) Area based results	Tropical	Woody	**0.86**	0.71	0.71	**0.86**	0.71	0.43	0.71	0.71
	Temperate	Woody	0.57	0.43	0.57	**0.71**	0.43	0.57	0.57	**0.71**
	Arid	Woody	**0.71**	0.43	**0.71**	**0.71**	0.43	0.43	**0.71**	**0.71**
	Cold	Woody	0.57	0.57	0.43	**0.71**	0.57	**0.71**	0.43	0.57
	Polar	Woody	0.43	**0.57**	0.29	0.14	0.29	0.29	0.14	0.43
	Tropical	Non-woody	0.43	0.29	0.43	0.43	**0.57**	0.43	0.14	0.43
	Temperate	Non-woody	0.57	0.57	0.71	**0.86**	0.57	**0.86**	0.43	0.57
	Arid	Non-woody	0.57	0.43	**0.71**	**0.71**	0.29	0.29	0.43	0.29
	Cold	Non-woody	0.86	0.71	0.71	**1**	0.57	0.43	0.57	0.57
	Polar	Non-woody	0.43	0.43	**0.57**	0.43	0.43	0.29	0.14	0.43

Table 7.9: Connectivity (i.e. Edge density) and modularity of plant trait networks for woody and non-woody species across five different climate regions. Analyses were ran using (A) mass-based, and (N) area based leaf N and P content.

Leaf nutrient content units		**(A) Mass based**		**(B) Area based**		
Group	**Climate**	**Present edges**	**Modularity**	**Present edges**	**Modularity**	n
Woody	Tropical	0.64	0	0.71	0	4414
Woody	Temperate	0.68	0.05	0.57	0.01	2695
Woody	Arid	0.64	0.06	0.61	0.15	1046
Woody	Cold	0.57	0.14	0.57	0.14	752
Woody	Polar	0.39	0.31	0.32	0.39	146
Non-woody	Tropical	0.32	0.27	0.39	0.1	537
Non-woody	Temperate	0.64	0	0.64	0.05	2352
Non-woody	Arid	0.43	0.2	0.46	0.19	873
Non-woody	Cold	0.64	0.05	0.68	0	2050
Non-woody	Polar	0.43	0.21	0.39	0.28	419

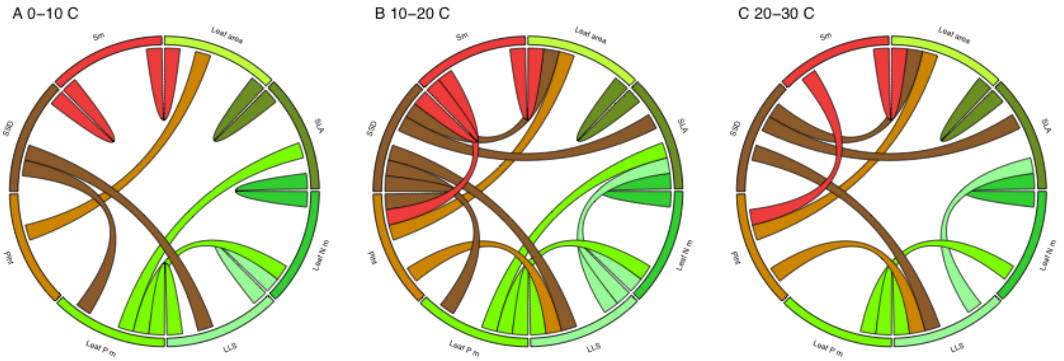


Figure 7.6: Connections among traits across multiple organs for woody angiosperms across a temperature gradient, holding precipitation between 0-500 mm.

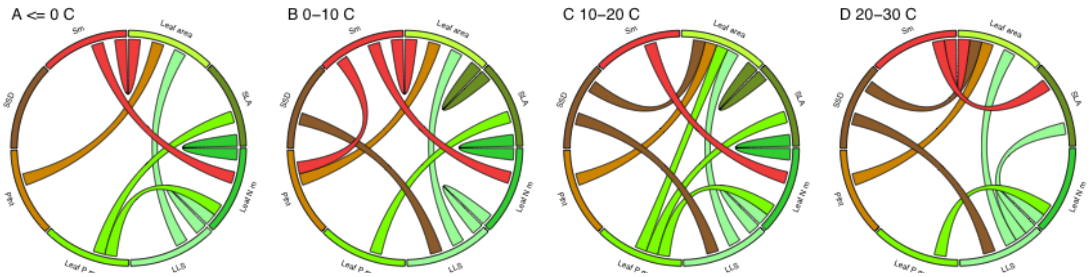


Figure 7.7: Connections among traits across multiple organs for forbs across a temperature gradient, holding precipitation constant between 0-500 mm.

Table 7.10: Number of present edges and modularity of plant trait networks and modularity for temperature and precipitation gradients. n refers to the number of species in each category.

Gradient	Group	Climate range	Present edges	Modularity	n
	Forb	0-500 mm	0.39	0.13	369
	Forb	500-1200 mm	0.43	0.14	916
Precipitation	Forb	1200-2500 mm	0.29	0.25	227
gradient	Monocots	0-500 mm	0.25	0.41	101
(Temperature	Monocots	500-1200 mm	0.29	0.3	348
10-20 degree	Monocots	1200-2500 mm	0.29	0.37	152
C)	Woody Angiosperms	0-500 mm	0.61	0.21	605
	Woody Angiosperms	500-1200 mm	0.57	0.25	1210
	Woody Angiosperms	1200-2500 mm	0.57	0.08	734
	Forb	0 C	0.29	0.22	166
	Forb	0-10 C	0.36	0.24	380
Temperature	Forb	10-20 C	0.39	0.13	369
gradient	Forb	20-30 C	0.32	0.17	156
(precipitation	Woody Angiosperms	0 C	0.29	0.07	NA
0-500 mm)	Woody Angiosperms	0-10 C	0.39	0.21	191
	Woody Angiosperms	10-20 C	0.61	0.21	605
	Woody Angiosperms	20-30 C	0.43	0.22	238

Table 7.11: Trait centrality (i.e. Degree) for woody angiosperms, non-woody forbs and non-woody monocots across a precipitation and temperature gradient, holding the other climate variable constant (see Section 7.4.6).

Gradient	Climate	Group	Leaf area	SLA	Leaf N	LLS	Leaf P	Plant height	SSD	Seed mass
	0-500 mm	Woody Angiosperms	0.57	0.71	0.43	0.71	0.57	0.57	0.86	0.43
	500-1200 mm	Woody Angiosperms	0.57	0.71	0.57	0.43	0.57	0.43	0.71	0.57
Precipitation	1200-2500 mm	Woody Angiosperms	0.57	0.71	0.43	0.71	0.43	0.43	0.57	0.71
gradient	0-500 mm	Forb	0.71	0.43	0.57	0.43	0.43	0.14	0.29	0.14
(Temperature	500-1200 mm	Forb	0.57	0.43	0.57	0.43	0.43	0.29	0.14	0.57
10-20 degree	1200-2500 mm	Forb	0.43	0.14	0.43	0.29	0.29	0.29	0.29	0.14
C)	0-500 mm	Mono	0.14	0.43	0.43	0.29	0.29	0.29	0	0.14
	500-1200 mm	Mono	0.14	0.29	0.43	0.57	0.29	0.29	0	0.29
	1200-2500 mm	Mono	0.29	0.14	0.29	0.29	0.29	0.43	0.14	0.43
	0 C	Woody Angiosperms	0.29	0.43	0.43	0.29	0.43	0.43	0	0
	0-10 C	Woody Angiosperms	0.43	0.43	0.43	0.43	0.57	0.14	0.43	0.29
Temperature	10-20 C	Woody Angiosperms	0.57	0.71	0.43	0.71	0.57	0.57	0.86	0.43
gradient	20-30 C	Woody Angiosperms	0.57	0.57	0.29	0.57	0.29	0.43	0.43	0.29
(precipitation	0 C	Forb	0.43	0.29	0.57	0.29	0.29	0.14	0	0.29
0-500 mm)	0-10 C	Forb	0.57	0.43	0.43	0.43	0.14	0.29	0.14	0.43
	10-20 C	Forb	0.71	0.43	0.57	0.43	0.43	0.14	0.29	0.14
	20-30 C	Forb	0.57	0.29	0.29	0.57	0.14	0.14	0.29	0.29

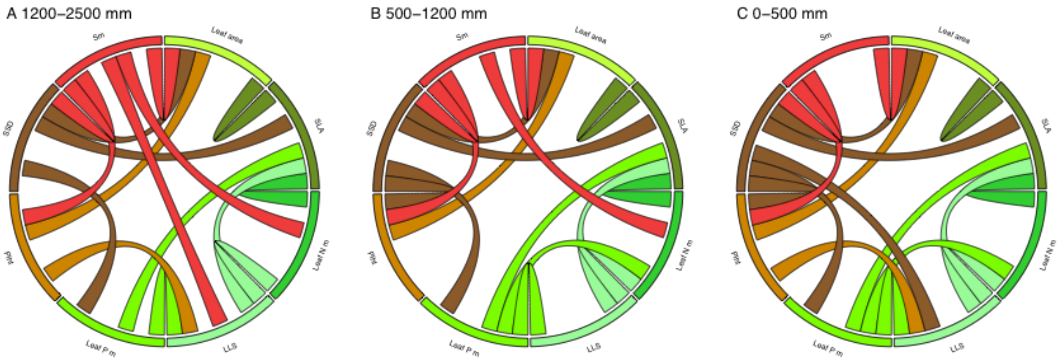


Figure 7.8: Connections among traits across multiple organs for woody angiosperms across a precipitation gradient, holding temperature between 10-20 C.

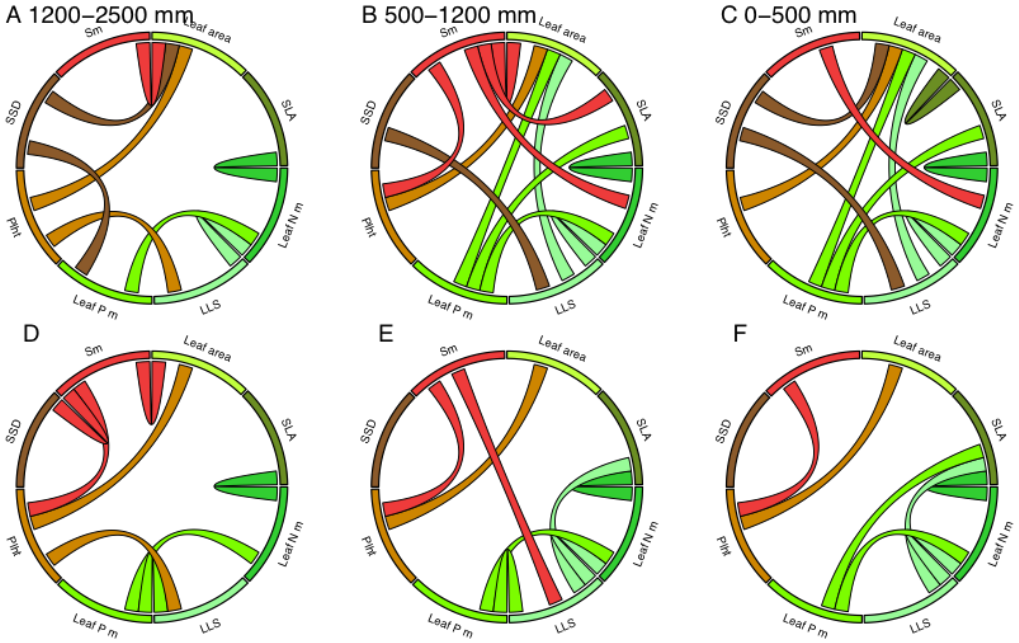
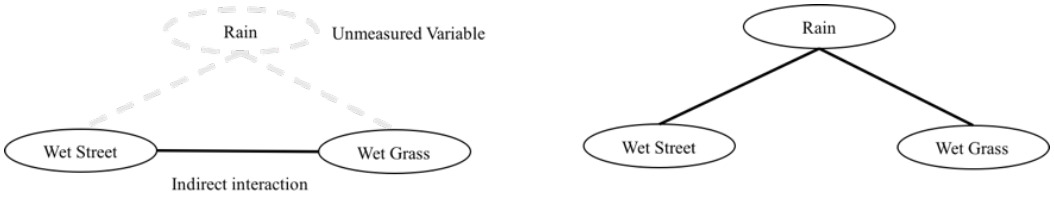
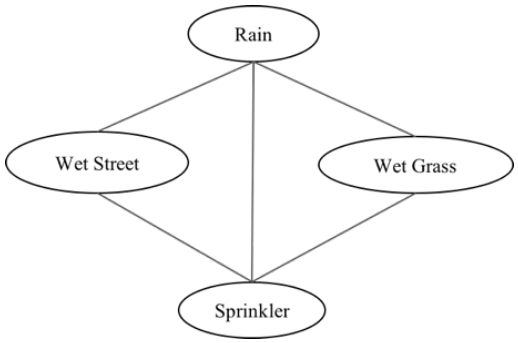


Figure 7.9: Connections among traits across multiple organs for non-woody (A-C) forbs and (D-E) monocots across a precipitation gradient, holding temperature constant between 10-20 C.



(a) Dependency graph without measuring Rain variable

(b) Dependency graph with measuring Rain Variable



(c) Dependency graph with measuring Sprinkler variable

Figure 7.10: Interaction between variables may depend on latent (or unmeasured) variables. For example, in graph a. Wet Street and Wet Grass are conditionally dependent without observing the Rain variable, but they become conditionally independent after observing the Rain Variable (b). Dash lines present the direct edges that cannot be obtained without considering the unmeasured variables. In another example (c), in the presence of a sprinkler the interaction structure may change to the following graph.

Chapter 8

Conclusion

In this thesis, we have presented our research on probabilistic structured models in high dimensional setting that advances understanding in three directions: theoretical advancements in structure learning of graphical models, advanced models in low rank matrix completion, and applications of probabilistic structured models in plant trait analysis.

In Chapter 3, we present the statistical analysis of direct change problem in Ising graphical models where any norm can be plugged in for characterizing the parameter structure. An optimization algorithm based on FISTA-style algorithms is proposed with the convergence rate of $O(1/T^2)$. We provide the statistical analysis for any norm such as L_1 norm, group sparse norm, node perturbation, etc. Our analysis is based on generic chaining and illustrates the important role of Gaussian widths (a geometric measure of size of suitable sets) in such results. For the special case of sparsity, we obtain a sharper result than previous results [132] under the same smooth density ratio assumption. Liu et al. [132] obtained the same result with a bounded density ratio assumption which is a more restrictive assumption. Although, we presented the results for Ising model, our analysis can be applied to any graphical model which satisfies the smooth density ratio assumption. Further, we extensively compared our generalized direct change estimator with an indirect approach over a wide range of graph structures and norms. We show that our direct approach has a better ROC curve than indirect approach without any assumption on the structure of individual graphs. We implemented indirect approach by estimating individual Ising model structures with L_1 norm regularizer. However,

if individual graphs has a suitable structure such as group sparsity, one may apply a regularization that can characterize the graph structure and may improve performance of the indirect approach. We will investigate this possibility in our future research.

In Chapter 4, we propose double plugin Gaussian (DoPinG) copula estimators to deal with non-Gaussian data with missing values. DoPinG estimates the sparse precision matrix corresponding to *non-paranormal* distributions by directly estimating non-parametric correlations, including Kendall's tau and Spearman's rho. DoPinG uses two plugin procedures, leveraging existing sparse precision estimators. We prove that DoPinG copula estimators consistently estimate the non-paranormal correlation matrix at a rate of $O\left(\frac{1}{(1-\delta)}\sqrt{\frac{\log p}{n}}\right)$, where δ is the probability of missing values. Through experiments we illustrate that by increasing number of missing values (increasing δ), the performance of the method get worse and the standard deviation is increasing in consistent with the theory. The performance of Kendall's tau and Spearman's rho is almost the same for the same percentage of missing values. Experimental results on non-Gaussian data show that DoPinG is significantly better than estimators like mGlasso, which are primarily designed for Gaussian data.

In Chapter 5, we studied the $\mathcal{MGIG}$ distribution and provided certain key properties with a novel sampling technique from the distribution. We show that the $\mathcal{MGIG}$ distribution is unimodal and the mode can be obtained by solving an ARE, and we propose a new importance sampling approach to infer the mean of $\mathcal{MGIG}$. The new sampler, chooses the proposal distribution to have the same mode as the $\mathcal{MGIG}$. This characterization leads to a far more effective sampler than [229, 233] since the new sampler align the shape of the proposal to the target distribution. Although, the $\mathcal{MGIG}$ has been recently applied to Bayesian models as the prior for the covariance matrix, here, we introduced a novel application of the $\mathcal{MGIG}$ in PMF or BPCA. We showed that the posterior distribution in PMF or BPCA has the $\mathcal{MGIG}$ distribution. This illustration, yields to a new CMC inference algorithm for PMF.

While the uncertainty quantification of a prediction is essential to understanding the prediction itself, most of the matrix completion methods give only a point estimate of missing entries without any uncertainty quantification. In Chapter 6, we show how we can derive uncertainty quantified estimates of missing values in sparse matrices. We propose BHPMF to incorporate the hierarchical side information and provide uncertainty

quantified estimates of the missing values. We developed a Gibbs sampling procedure for inference in the model. We observe that block-wise sampling with diagonal covariance as traits' prior outperforms point-wise sampling, which uses a full covariance trait structure as prior over traits. BHPMF with block-wise sampling provides higher point estimation accuracy than PMF, HPMF, which is the state-of-the-art for trait prediction, and MEAN, which is frequently used as a baseline in the ecology community. We then generalized BHPMF to consider hierarchical multiple inheritance side information (MI-BHPMF). We show that the Gibbs sampler readily generalizes to this setting. We illustrate that BHPMF and MI-BHPMF are accurate (small RMSE) when they are confident (small standard deviation), whereas the error is high when the uncertainty is high. On the example of 13 plant traits from the world's largest plant trait database (TRY) and the MovieLens data set we demonstrate that this hypothesis is fulfilled in all cases. Quantified uncertainty estimates based on BHPMF and MI-BHPMF thus help to identify areas of limited confidence, which can be used to inform future trait data collection efforts.

Lastly, in Chapter 7, we applied the sparse structure of graphical model estimators to learn the plant trait-trait interactions. Moving from Tropical to Polar environments, the density of connections among traits in woody plants decreases, while decoupling across organs (i.e. modularity) increases. We found no clear pattern across these climate types for non-woody species. Further analyses revealed that for Monocots and woody Angiosperms the density of connections and modularity increases with decreasing precipitation, Forbs showed the opposite pattern in modularity. Increase in temperature was related to increasing in modularity but only for woody angiosperms, with inconsistent patterns for all other groups. No trait was invariably central across climate regions, temperature or precipitation gradients. Connectivity among plant traits is independent of the strength of trait-trait correlations or trade-offs. These analyses show that connections and coupling among traits is dependent on local environmental conditions, and differ across growth forms. A fully integrated plant strategy may not be advantageous under all environmental conditions. The way in which specific combinations of traits may influence the success of species in a given environment change across climate gradients and growth forms.

References

[1] P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research*, 7:1743–1788, 2006.

[2] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup. Scalable tensor factorizations with missing data. In *SDM*, 2010.

[3] D. Ackerly. Functional strategies of chaparral shrubs in relation to seasonal water deficit and disturbance. *Ecological Monographs*, 74(1):25–44, 2004.

[4] D. Aggarwal and S. Merugu. Predictive discrete latent factor models for large scale dyadic data. In *KDD*, 2007.

[5] A. Agovic, A. Banerjee, and S. Chatterjee. Probabilistic Matrix Addition. In *ICML*, 2011.

[6] C. H. Albert, W. Thuiller, N. G. Yoccoz, A. Soudant, F. Boucher, P. Saccone, and S. Lavorel. Intraspecific functional variability: extent, structure and sources of variation. *Journal of Ecology*, 98(3):604–613, 2010.

[7] U. Alon. Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867, 2003.

[8] A. Anandkumar, V. Tan, and A. S. Willsky. High-dimensional graphical model selection: tractable graph families and necessary conditions. In *Advances in Neural Information Processing Systems*, pages 1863–1871, 2011.

[9] B. D. Anderson and J. B. Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[10] G. Andrew and J. Gao. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM, 2007.

[11] Z. Bai and J. Demmel. Using the matrix sign function to compute invariant subspaces. *SIAM Journal on Matrix Analysis and Applications*, 19(1):205–225, 1998.

[12] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach*. MIT press, 2001.

[13] N. M. Ball and R. J. Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, 2010.

[14] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with Norm Regularization. In *Neural Information Processing Systems*, 2014.

[15] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[16] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[17] C. Baraloto, C. E. T. Paine, et al. Functional trait variation and sampling strategies in species-rich plant communities. *Functional Ecology*, 24(1):208–216, 2009.

[18] C. Baraloto, C. Timothy Paine, L. Poorter, J. Beauchene, D. Bonal, A.-M. Domenach, B. Hérault, S. Patino, J.-C. Roggy, and J. Chave. Decoupled leaf and stem economics in rain forest trees. *Ecology letters*, 13(11):1338–1347, 2010.

[19] O. Barndorff-Nielsen, P. Blæsild, et al. Exponential transformation models. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 379(1776):41–65, 1982.

[20] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[21] C. M. Bishop. Bayesian PCA. *NIPS*, pages 382–388, 1999.

[22] C. M. Bishop. Variational principal components. In *ICANN*, 1999.

[23] D. Blei, P. Cook, and M. Hoffman. Bayesian nonparametric matrix factorization for recorded music. In *ICML*, pages 439–446, 2010.

[24] A. J. Bloom, F. S. Chapin, and H. A. Mooney. Resource limitation in plants–an economic analogy. *Annual review of Ecology and Systematics*, pages 363–392, 1985.

[25] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[26] C. K. Boyce, T. J. Brodribb, T. S. Feild, and M. A. Zwieniecki. Angiosperm leaf vein evolution was physiologically and environmentally transformative. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1663):1771–1776, 2009.

[27] S. Boyd, E. C. N. Parikh, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends Machine Learning*, 3(1), 2011.

[28] S. P. Boyd and C. H. Barratt. *Linear controller design: limits of performance*. 1991.

[29] A. D. Brevern, S. Hazout, and A. Malpertuy. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC bioinformatics*, 5(1):114, 2004.

[30] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[31] A. Bunse-Gerstner and V. Mehrmann. A symplectic qr like algorithm for the solution of the real algebraic riccati equation. *Automatic Control, IEEE Transactions on*, 31(12):1104–1113, 1986.

[32] R. W. Butler. Generalized inverse Gaussian distributions and their Wishart connections. *Scandinavian journal of statistics*, 25(1):69–75, 1998.

[33] R. W. Butler and A. Wood. Laplace approximation for bessel functions of matrix argument. *Journal of Computational and Applied Mathematics*, 155(2):359–382, 2003.

[34] R. Byers. Solving the algebraic Riccati equation with the matrix sign function. *Linear Algebra and its Applications*, 85:267–279, 1987.

[35] T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 2012.

[36] T. Cai, C. Zhang, and H. Zhou. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *American Statistical Association*, 106:594–607, 2011.

[37] T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, page ass058, 2012.

[38] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[39] G. C. Cawley and N. L. Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355, 2006.

[40] L. Cayuela. Taxonstand: Taxonomic standardization of plant species names. *R package version*, 1, 2014.

[41] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

[42] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[43] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly. Sparse Group Lasso: Consistency and Climate Applications. *SIAM International Conference on Data Mining (SDM)*, 2012.

[44] J. Chave, D. Coomes, et al. Towards a worldwide wood economics spectrum. *Ecology Letters*, 12(4):351–366, 2009.

[45] J. Chave, D. Coomes, S. Jansen, S. L. Lewis, N. G. Swenson, and A. E. Zanne. Towards a worldwide wood economics spectrum. *Ecology letters*, 12(4):351–366, 2009.

[46] S. Chen and A. Banerjee. Structured estimation with atomic norms: General bounds and applications. In *Neural Information Processing Systems*, 2015.

[47] S. Chen and D. Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

[48] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[49] Y. Chen, Y. Gu, and A. O. Hero. Sparse LMS for System Identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[50] J. Cheng, P. Chu, D. Chen, and Y. Bai. Functional correlations between specific leaf area and specific root length along a regional environmental gradient in inner mongolia grasslands. *Functional Ecology*, 2015.

[51] D. M. Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[52] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

[53] J. S. Clark. Individuals and the variation needed for high species diversity in forest trees. *Science*, 327(5969):1129–1132, 2010.

[54] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[55] J. Cornelissen. A triangular relationship between leaf size and seed size among woody species: allometry, ontogeny, ecology and taxonomy. *Oecologia*, 118(2):248–255, 1999.

[56] J. Craine, D. Tilman, D. Wedin, P. Reich, M. Tjoelker, and J. Knops. Functional traits, productivity and effects on nitrogen cycling of 33 grassland species. *Functional Ecology*, 16(5):563–574, 2002.

[57] G. R. Cross and A. K. Jain. Markov random field texture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1):25–39, 1983.

[58] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[59] D. Das. *Bayesian sparse regression with application to data-driven understanding of climate*. TEMPLE UNIVERSITY, 2015.

[60] S. Dasgupta. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141. Morgan Kaufmann Publishers Inc., 1999.

[61] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–668, 2005.

[62] E. G. de la Riva, A. Tosto, I. M. Pérez-Ramos, C. M. Navarro-Fernández, M. Olmo, N. P. Anten, T. Marañón, and R. Villar. A plant economics spectrum in mediterranean forests along environmental gradients: is there coordination among leaf, stem and root traits? *Journal of Vegetation Science*, 27(1):187–199, 2016.

[63] S. Díaz, J. G. Hodgson, K. Thompson, et al. The plant traits that drive ecosystems: evidence from three continents. *Journal of Vegetation Science*, 15(3):295–304, 2004.

[64] S. Díaz, J. Kattge, J. H. Cornelissen, I. J. Wright, S. Lavorel, S. Dray, B. Reu, M. Kleyer, C. Wirth, I. C. Prentice, et al. The global spectrum of plant form and function. *Nature*, 529(7585):167–171, 2016.

[65] R. Dybzinski, C. Farrior, A. Wolf, P. B. Reich, and S. W. Pacala. Evolutionarily stable strategy carbon allocation to foliage, wood, and fine roots in trees competing for light and nitrogen: an analytically tractable, individual-based model and quantitative comparisons to data. *The American Naturalist*, 177(2):153–166, 2011.

[66] E. Eberlein and U. Keller. Hyperbolic distributions in finance. *Bernoulli*, pages 281–299, 1995.

[67] I. Ebert-Uphoff and Y. Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

[68] H. Fang, K. Fang, and S. Kotz. The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82:1–16, 2002.

[69] F. Fazayeli and A. Banerjee. Generalized direct change estimation in ising model structure. *ICML*, 2016.

[70] F. Fazayeli and A. Banerjee. The matrix generalized inverse gaussian distribution: Properties and applications. *ECML-PKDD*, 2016.

[71] F. Fazayeli, A. Banerjee, J. Kattge, F. Schrodt, and P. Reich. Uncertainty quantified matrix completion using bayesian hierarchical matrix factorization. In *ICMLA*, 2014.

[72] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *ACC*, volume 6, pages 4734–4739. IEEE, 2001.

[73] H. Flores-Moreno, F. Fazayeli, A. Banerjee, A. Datta, J. Kattge, E. E. Butler, O. K. Atkin, K. Wythers, M. Chen, M. Anand, M. Bahn, S. Burrascano, C. Byun, J. H. C. Cornelissen, J. Craine, A. Gonzalez-Melo, W. N. Hattingh, S. Jansen, N. J. Kraft, K. Kramer, D. C. Laughlin, V. Minden, U. Niinemets, V. Onipchenko, J. Peñuelas, N. A. Soudzilovskaia, and P. B. Reich. Robustness of trait connections between multiple plant organs across environmental gradients and growth forms. *submitted*, 2017.

[74] C. Fortunel, P. V. Fine, and C. Baraloto. Leaf, stem and root tissue strategies across 758 neotropical tree species. *Functional Ecology*, 26(5):1153–1161, 2012.

[75] G. T. Freschet, J. H. Cornelissen, R. S. Van Logtestijn, and R. Aerts. Evidence of the plant economics spectrumin a subarctic flora. *Journal of Ecology*, 98(2):362–373, 2010.

[76] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[77] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[78] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[79] Q. Fu and A. Banerjee. Multiplicative mixture models for overlapping clustering. In *ICDM*, pages 791–796, 2008.

[80] J. Gao, G. Andrew, M. Johnson, and K. Toutanova. A comparative study of parameter estimation methods for statistical natural language processing. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 824, 2007.

[81] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[82] P. Giudici and A. Spelta. Graphical network models for international financial flows. *Journal of Business & Economic Statistics*, 34(1):128–138, 2016.

[83] Y. Gordon. On Milman's Inequality and Random Subspaces Which Escape Through a Mesh in Rn. In *Geometric Aspects of Functional Analysis*, volume 1317 of *Lecture Notes in Mathematics*, pages 84–106. Springer Berlin, 1988.

[84] C. W. Granger. Seasonality: causation, interpretation, and implications. In *Seasonal analysis of economic time series*, pages 33–56. NBER, 1979.

[85] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[86] K. Heller and Z. Ghahramani. A nonparametric bayesian approach to modeling overlapping clusters. 2007.

[87] C. S. Herz. Bessel functions of matrix argument. *Annals of Mathematics*, pages 474–523, 1955.

[88] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[89] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[90] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

[91] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[92] S. Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.

[93] C. Hu, L. Cheng, J. Sepulcre, K. A. Johnson, G. E. Fakhri, Y. M. Lu, and Q. Li. A spectral graph regression model for learning brain connectivity of alzheimers disease. *PloS one*, 10(5):e0128136, 2015.

[94] A. Ishida, T. Nakano, K. Yazaki, S. Matsuki, N. Koike, D. L. Lauenstein, M. Shimizu, and N. Yamashita. Coordination between leaf and stem traits related to leaf carbon gain and hydraulics across 32 drought-tolerant angiosperms. *Oecologia*, 156(1):193–202, 2008.

[95] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[96] M. M. Jager, S. J. Richardson, P. J. Bellingham, M. J. Clearwater, and D. C. Laughlin. Soil fertility induces coordinated responses of multiple independent functional traits. *Journal of Ecology*, 103(2):374–385, 2015.

[97] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. *Symposium on Theory of Computing*, 2013.

[98] J. Jankova, S. van de Geer, et al. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[99] B. Jørgensen. *Statistical properties of the generalized inverse Gaussian distribution*. Springer, 1982.

[100] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[101] J. Kattge, S. Diaz, S. Lavorel, et al. Try–a global database of plant traits. *Global Change Biology*, 17(9):2905–2935, 2011.

[102] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):606–617, 2007.

[103] M. Kolar and E. Xing. Estimating sparse precision matrices from data with missing values. In *ICML*, 2012.

[104] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear Norm Penalization and Optimal Rates for Noisy Low Rank Matrix Completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[105] A. Kong, J. Liu, and W. Wong. Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288, 1994.

[106] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer*, 2009.

[107] D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2:193, 2008.

[108] N. J. Kraft, O. Godoy, and J. M. Levine. Plant functional traits and the multidimensional nature of species coexistence. *Proceedings of the National Academy of Sciences*, 112(3):797–802, 2015.

[109] K. R. Kramer-Walter, P. J. Bellingham, T. R. Millar, R. D. Smissen, S. J. Richardson, and D. C. Laughlin. Root traits are multidimensional: specific root length is independent from root tissue density and the plant economic spectrum. *Journal of Ecology*, 2016.

[110] W. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[111] G. Kunstler, D. Falster, D. A. Coomes, F. Hui, R. M. Kooyman, D. C. Laughlin, L. Poorter, M. Vanderwel, G. Vieilledent, S. J. Wright, et al. Plant functional traits have globally consistent effects on competition. *Nature*, 529(7585):204–207, 2016.

[112] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[113] W. Larcher. *Physiological plant ecology: ecophysiology and stress physiology of functional groups*. Springer Science & Business Media, 2003.

[114] A. J. Laub. A schur method for solving algebraic riccati equations. *Automatic Control, IEEE Transactions on*, 24(6):913–921, 1979.

[115] D. C. Laughlin, J. J. Leppert, M. M. Moore, and C. H. Sieg. A multi-trait test of the leaf-height-seed plant strategy scheme with 133 species from a pine forest flora. *Functional Ecology*, 24(3):493–501, 2010.

[116] N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

[117] N. Lawrence and R. Urtasun. Non-linear Matrix Factorization with Gaussian Processes. In *ICML*, 2009.

[118] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2013.

[119] P. Legendre and L. F. Legendre. *Numerical ecology*, volume 24. Elsevier, 2012.

[120] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[121] F. L. Li and W. K. Bao. New insights into leaf and fine-root trait relationships: implications of resource acquisition among 23 xerophytic woody species. *Ecology and Evolution*, 5(22):5344–5351, 2015.

[122] H. Li and J. Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.

[123] T. Li, E. Chu, W. Lin, and P. Weng. Solving large-scale continuous-time algebraic riccati equations by doubling. *Journal of Computational and Applied Mathematics*, 237(1):373–383, 2013.

[124] X. Li, T. Zhao, and H. Liu. camel: Calibrated machine learning. *R package version 0.2. 0*, 2013.

[125] H. Lipson, J. B. Pollack, N. P. Suh, and P. Wainwright. On the origin of modular variation. *Evolution*, 56(8):1549–1556, 2002.

[126] R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1987.

[127] G. Liu, G. T. Freschet, X. Pan, J. H. Cornelissen, Y. Li, and M. Dong. Coordinated variation in leaf and root traits across multiple spatial scales in chinese semi-arid and arid ecosystems. *New Phytologist*, 188(2):543–553, 2010.

[128] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(40):2293–2326, 2012.

[129] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. 10:2295–2328, 2009.

[130] H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Preprint*, 2012.

[131] S. Liu, J. A. Quinn, M. U. Gutmann, T. Suzuki, and M. Sugiyama. Direct learning of sparse changes in markov networks by density ratio estimation. *Neural computation*, 26(6):1169–1197, 2014.

[132] S. Liu, T. Suzuki, and M. Sugiyama. Support consistency of direct sparse-change learning in markov networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[133] S. Liu, T. Suzuki, and M. Sugiyama. Support consistency of direct sparse-change learning in markov networks. In *arXiv:1407.0581v10*, 2015.

[134] P. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *NIPS*, 2012.

[135] P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[136] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *ArXiv*, 2012.

[137] D. MacKay. *Information theory, inference, and learning algorithms*, volume 7. Citeseer, 2003.

[138] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. 1999.

[139] B. M. Marlin, M. Schmidt, and K. P. Murphy. Group sparse priors for covariance estimation. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 383–392. AUAI Press, 2009.

[140] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[141] A. K. Menon, K. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *KDD*, pages 141–149. ACM, 2011.

[142] J. Messier, B. J. McGill, B. J. Enquist, and M. J. Lechowicz. Trait variation and integration across scales: is the leaf economic spectrum present at local scales? *Ecography*, 2016.

[143] D. Meziane and B. Shipley. Direct and indirect relationships between specific leaf area, leaf nitrogen and leaf gas exchange. effects of irradiance and nutrient supply. *Annals of Botany*, 88(5):915–927, 2001.

[144] T. P. Minka. Automatic choice of dimensionality for pca. In *NIPS*, volume 13, pages 598–604, 2000.

[145] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

[146] G. Mohler. Learning convolution filters for inverse covariance estimation of neural network connectivity. In *Advances in Neural Information Processing Systems*, pages 891–899, 2014.

[147] A. T. Moles, D. D. Ackerly, C. O. Webb, J. C. Tweddle, J. B. Dickie, and M. Westoby. A brief history of seed size. *Science*, 307(5709):576–580, 2005.

[148] A. T. Moles, D. I. Warton, L. Warman, N. G. Swenson, S. W. Laffan, A. E. Zanne, A. Pitman, F. A. Hemmings, and M. R. Leishman. Global patterns in plant height. *Journal of Ecology*, 97(5):923–932, 2009.

[149] A. T. Moles and M. Westoby. Seed size and plant strategy across the whole life cycle. *Oikos*, 113(1):91–105, 2006.

[150] S. Negahban and M. J. Wainwright. Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise. *Journal of Machine Learning Research (JMLR)*, 13(1):1665–1697, 2012.

[151] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[152] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[153] Ü. Niinemets. A review of light interception in plant stands from leaf to canopy in different plant functional types and in species with varying shade tolerance. *Ecological Research*, 25(4):693–714, 2010.

[154] K. J. Niklas. *Plant allometry: the scaling of form and process*. University of Chicago Press, 1994.

[155] I. R. Noble and H. Gitay. A functional classification for predicting the dynamics of landscapes. *Journal of Vegetation science*, 7(3):329–336, 1996.

[156] J. L. Osnas, J. W. Lichstein, P. B. Reich, and S. W. Pacala. Global leaf trait relationships: mass, area, and the leaf economics spectrum. *Science*, 340(6133):741–744, 2013.

[157] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[158] M. C. Peel, B. L. Finlayson, and T. A. McMahon. Updated world map of the köppen-geiger climate classification. *Hydrology and earth system sciences discussions*, 4(2):439–473, 2007.

[159] M. W. Pennell, R. G. FitzJohn, and W. K. Cornwell. A simple approach for maximizing the overlap of phylogenetic and comparative data. *Methods in Ecology and Evolution*, 2016.

[160] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84(9):2347–2363, 2003.

[161] H. Poorter, H. Lambers, and J. R. Evans. Trait correlation networks: a whole-plant perspective on the recently criticized leaf economic spectrum. *New Phytologist*, 201(2):378–382, 2014.

[162] I. Porteous, A. Asuncion, and M. Welling. Bayesian matrix factorization with side information and dirichlet process mixtures. In *AAAI*, 2010.

[163] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[164] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[165] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[166] B. Recht. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research (JMLR)*, 12:3413–3430, 2011.

[167] P. Reich, M. Walters, and D. Ellsworth. Leaf life-span in relation to leaf, plant, and stand characteristics among diverse ecosystems. *Ecological monographs*, 62(3):365–392, 1992.

[168] P. B. Reich. The world-wide fast–slowplant economics spectrum: a traits manifesto. *Journal of Ecology*, 102(2):275–301, 2014.

[169] P. B. Reich, C. Buschena, M. Tjoelker, K. Wrage, J. Knops, D. Tilman, and J. Machado. Variation in growth rate and ecophysiology among 34 grassland and savanna species under contrasting n supply: a test of functional group differences. *New Phytologist*, 157(3):617–631, 2003.

[170] P. B. Reich, D. S. Ellsworth, M. B. Walters, J. M. Vose, C. Gresham, J. C. Volin, and W. D. Bowman. Generality of leaf trait relationships: a test across six biomes. *Ecology*, 80(6):1955–1969, 1999.

[171] P. B. Reich, M. B. Walters, and D. S. Ellsworth. From tropics to tundra: global convergence in plant functioning. *Proceedings of the National Academy of Sciences*, 94(25):13730–13734, 1997.

[172] B. D. Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2005.

[173] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[174] R. Salakhutdinov and A. Mnih. Probabilistic Matrix Factorization. In *NIPS*, 2007.

[175] R. Salakhutdinov and A. Mnih. Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. In *ICML*, 2008.

[176] R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *NIPS*, 2010.

[177] H. J. Schenk, S. Espino, C. M. Goedhart, M. Nordenstahl, H. I. M. Cabrera, and C. S. Jones. Hydraulic integration and shrub growth form linked across continental aridity gradients. *Proceedings of the National Academy of Sciences*, 105(32):11248–11253, 2008.

[178] J. T. Schoof and S. Pryor. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of climatology*, 21(7):773–790, 2001.

[179] F. Schrodt, J. Kattge, H. Shan, F. Fazayeli, J. Joswig, A. Banerjee, M. Reichstein, G. Bönisch, S. Díaz, J. Dickie, et al. Bhpmf–a hierarchical bayesian approach

to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, 2015.

[180] V. Seshadri. Some properties of the matrix generalized inverse Gaussian distribution. *Statistical methods and practice. Recent advances. Narosa Publishing House, New Delhi*, pages 47–56, 2003.

[181] V. Seshadri and J. Wesołowski. More on connections between Wishart and matrix GIG distributions. *Metrika*, 68(2):219–232, 2008.

[182] H. Shan and A. Banerjee. Generalized Probabilistic Matrix Factorizations for Collaborative Filtering. In *ICDM*, 2010.

[183] H. Shan, J. Kattge, P. B. Reich, A. Banerjee, F. Schrodt, and M. Reichstein. Gap filling in the plant kingdom—trait prediction using hierarchical probabilistic matrix factorization. *ICML*, 2012.

[184] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[185] B. Shipley, M. J. Lechowicz, I. Wright, and P. B. Reich. Fundamental trade-offs generating the worldwide leaf economics spectrum. *Ecology*, 87(3):535–541, 2006.

[186] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[187] Y. Singer and J. C. Duchi. Efficient learning using forward-backward splitting. In *Neural Information Processing Systems*, pages 495–503, 2009.

[188] A. Singh and G. Gordon. A Bayesian Matrix Factorization Model for Relational Data. In *UAI*, 2010.

[189] W. Smith and R. Hocking. Algorithm as 53: Wishart variate generator. *Applied Statistics*, 1972.

[190] N. Srebro. Maximum likelihood bounded tree-width markov networks. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 504–511. Morgan Kaufmann Publishers Inc., 2001.

[191] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-Margin Matrix Factorization. In *NIPS*, 2005.

[192] N. Stadler and P. Buhlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, pages 1–17, 2009.

[193] J.-L. Starck, D. L. Donoho, and E. J. Candès. Astronomical image representation by the curvelet transform. *Astronomy & Astrophysics*, 398(2):785–800, 2003.

[194] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

[195] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[196] I. Sutskever, R. Salakhutdinov, and J. Tenenbaum. Modelling Relational Data using Bayesian Clustered Tensor Facotrization. In *NIPS*, 2009.

[197] M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, pages 1049–1103, 1996.

[198] M. Talagrand. Majorizing measures without measures. *Annals of probability*, pages 411–417, 2001.

[199] M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer Berlin, 2005.

[200] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

[201] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[202] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999.

[203] M. Tjoelker, J. Craine, D. Wedin, P. Reich, and D. Tilman. Linking leaf and root trait syndromes among 39 grassland and savannah species. *New Phytologist*, 167(2):493–508, 2005.

[204] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[205] H. Tsukahara. Efficient estimation in the bivariate normal copula model: Normal margins are least-favorable. *Bernoulli*, 3:55–77, 1997.

[206] H. Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33:357–375, 2005.

[207] J. C. Uyeda, D. S. Caetano, and M. W. Pennell. Comparative analysis of principal components can be misleading. *Systematic biology*, page syv019, 2015.

[208] V. Vapnik and R. Izmailov. Statistical inference problems and their rigorous solutions. In *Statistical Learning and Data Sciences*, pages 33–71. Springer International Publishing, 2015.

[209] R. Vershynin. Estimation in high dimensions: a geometric perspective. *Sampling Theory, a Renaissance*, 2014.

[210] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2201, 2009.

[211] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD*, pages 448–456. ACM, 2011.

[212] H. Wang and A. Banerjee. Online alternating direction method. *ICML*, 2012.

[213] H. Wang, A. Banerjee, C. Hsieh, P. Ravikumar, and I. Dhillon. Large scale distributed sparse precision estimation. In *NIPS*, 2013.

[214] H. Wang, F. Fazayeli, S. Chatterjee, A. Banerjee, K. Steinhauser, A. Ganguly, K. Bhattacharjee, A. Konar, and A. Nagar. Gaussian copula precision estimation with missing values. In *AISTATS*, pages 978–986, 2014.

[215] Y. Wang, X. Lu, I. Wright, Y. Dai, P. Rayner, and P. Reich. Correlations among leaf traits provide a significant constraint on the estimate of global gross primary production. *Geophysical Research Letters*, 39(19), 2012.

[216] D. Welsh. *Complexity: knots, colourings and countings*, volume 186. Cambridge university press, 1993.

[217] M. Westoby, D. S. Falster, A. T. Moles, P. A. Vesk, and I. J. Wright. Plant ecological strategies: some leading dimensions of variation between species. *Annual review of ecology and systematics*, 33(1):125–159, 2002.

[218] M. Westoby, M. R. Leishman, and J. M. Lord. On misinterpreting thephylogenetic correction'. *Journal of Ecology*, 83(3):531–534, 1995.

[219] M. Westoby, P. B. Reich, and I. J. Wright. Understanding ecological variation across species: area-based vs mass-based expression of leaf traits. *New Phytologist*, 199(2):322–323, 2013.

[220] M. Westoby and I. J. Wright. The leaf size–twig size spectrum and its relationship to other important spectra of variation among species. *Oecologia*, 135(4):621–628, 2003.

[221] R. H. Whittaker et al. Communities and ecosystems. *Communities and ecosystems.*, 1970.

[222] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

[223] I. J. Wright, D. D. Ackerly, F. Bongers, K. E. Harms, G. Ibarra-Manriquez, M. Martinez-Ramos, S. J. Mazer, H. C. Muller-Landau, H. Paz, N. C. Pitman, et al. Relationships among ecologically important dimensions of plant trait variation in seven neotropical forests. *Annals of Botany*, 99(5):1003–1015, 2007.

[224] I. J. Wright, P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, T. Chapin, J. H. Cornelissen, M. Diemer, et al. The worldwide leaf economics spectrum. *Nature*, 428(6985):821–827, 2004.

[225] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Journal of the ACM*, 2009.

[226] L. Xiong, X. Chen, T. Huang, J. G. Schneider, and J. G. Carbonell. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *SDM*, 2010.

[227] L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

[228] E. Yang, A. Genevera, Z. Liu, and P. K. Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

[229] M. Yang, Y. Li, and Z. Zhang. Multi-task learning with Gaussian matrix generalized inverse Gaussian model. In *ICML*, 2013.

[230] Z. Yang, R. Algesheimer, and C. J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6, 2016.

[231] J. Ye and J. Liu. Sparse Methods for Biomedical Data. *ACM SIGKDD Explorations Newsletter*, 14(1):4–15, 2012.

[232] M. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002.

[233] K. Yoshii and R. Tomioka. Infinite positive semidefinite tensor factorization for source separation of mixture signals. In *ICML*, 2013.

[234] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pages 352–360, 2011.

[235] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *JMLR*, 11, 2010.

[236] M. Yuan and Y. Lin. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(1):49–67, 2007.

[237] A. E. Zanne, M. Westoby, D. S. Falster, D. D. Ackerly, S. R. Loarie, S. E. Arnold, and D. A. Coomes. Angiosperm wood structure: global patterns in vessel anatomy and their relation to wood density and potential conductivity. *American Journal of Botany*, 97(2):207–215, 2010.

[238] S. Zhao, T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

[239] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling Disease Progression via Fused Sparse Group Lasso. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[240] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling Disease Progression via Multi-Task Learning. 78:233–248, 2013.

[241] T. Zhou, H. Shan, A. Banerjee, et al. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, 2012.

[242] D. Zimmerman, B. Zumbo, and R. Williams. Bias in estimation and hypothesis testing of correlation. *Transformation*, 24:133–158, 2003.

Appendix A

Direct Change Estimation Appendix

A.1 Background and Preliminaries

Definition A.1.1 Sub-Gaussian random variable: *We say that a random variable x is sub-Gaussian if the moments satisfies*

$$[E|x|^p]^{\frac{1}{p}} \leq K_2\sqrt{p} \tag{A.1}$$

for any $p \geq 1$ with a constant K_2. The minimum value of K_2 is called sub-Gaussian norm of x, denoted by $|||x|||_{\psi_2}$. If $E[x] = 0$, then

$$E[\exp\{tX\}] \leq \exp\{Ct^2|||X|||_{\Psi_2}^2\}, \tag{A.2}$$

where C and c are positive constant.

Definition A.1.2 Sub-Gaussian random vector: *We say that a random vector X in $\mathbb{R}^n$ is sub-Gaussian if the one-dimensional marginals $\langle X, \mathbf{x}\rangle$ are sub-Gaussian random variables for all $\mathbf{x} \in \mathbb{R}^n$. The sub-Gaussian norm of X is defined as*

$$|||X|||_{\psi_2} = \sup_{\mathbf{x}\in S^{n-1}} \|\langle X, \mathbf{x}\rangle\|_{\psi_2} \tag{A.3}$$

Lemma 17 *Consider a sub-Gaussian vector $X \in \mathbb{R}^n$ with $\|\|X\|\|_{\Psi_2} < K$, then for any vector u, $\langle X, u\rangle$ is a sub-Gaussian variable with $\|\|\langle X, u\rangle\|\| < K\|u\|_2$.*

Proof: The argument is based on Definition A.1.2 as follows,

$$\|\|\langle X, u\rangle\|\|_{\Psi_2} = \|u\|_2 \left\|\left\| \left\langle X, \frac{u}{\|u\|_2} \right\rangle \right\|\right\|_{\Psi_2} \leq \|u\|_2 \sup_{x \in S^{n-1}} \langle X, x\rangle = \|u\|_2 \|\|X\|\|_{\Psi_2} \leq K\|u\|_2. \tag{A.4}$$

■

Lemma 18 *Let X_1 and X_2 be centered sub-Gaussian random variables with $\|\|X_1\|\|_{\Psi_2} = b_1$ and $\|\|X_2\|\|_{\Psi_1} = b_2$. Then $X_1 + X_2$ is centered sub-Gaussian with $\|\|X_1 + X_2\|\|_{\Psi_2} = b_1 + b_2$.*

Proof: The argument is based on the definition of moment generating function of sub-Gaussian random variable:

Using Holder inequality for any $p, q > 0$ where $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$\begin{aligned} E[\exp\{t(X_1 + X_2)\}] &\leq (E[\exp\{tX_1\}^p])^{1/p} (E[\exp\{tX_1\}^q])^{1/q} \\ &\leq \exp\{Ct^2(pb_1^2 + qb_2^2)\} = \exp\{Ct^2(pb_1^2 + \frac{p}{1-p} b_2^2)\}. \end{aligned} \tag{A.5}$$

The minimum of (A.5) occurs with $p = \frac{b_2}{b_1}$. As a result, we have

$$E[\exp\{t(X_1 + X_2)\}] \leq \exp\{Ct^2(b_1 + b_2)^2\}. \tag{A.6}$$

The proof is complete. ■

A.1.1 Generic Chaining

Definition A.1.3 (Majorizing measure [197]) *Given $\alpha > 0$, and a metric space (T, d) (that need not be finite), we define*

$$\gamma_\alpha(T, d) = \inf \sup_t \sum_{n \geq 0} 2^{n/\alpha} \Delta(A_n(T)). \tag{A.7}$$

where the infimum is taken over all admissible sequences and $\Delta(A_n(T))$ is the diameter of $A_n(t)$.

Note that $\gamma_2(T, \|.\|_2)$ coincides with the Gaussian width of T: $w(T)$.

Lemma 19 *Given a metric space (T, d), we have*

$$\gamma_1(T, \|.\|_\infty) \leq \gamma_2^2(T, \|.\|_2). \tag{A.8}$$

Proof: Define $d_2(s,t) = \|s - t\|_2$ and $d_1(s,t) = \|s - t\|_\infty$. We use the traditional definition of majorizing measure $\gamma_{\alpha,1}(T, d)$ from [197]

$$\gamma_{\alpha,1}(T, d) = \inf \sup_t \left(\int_0^\infty \left(\log \frac{1}{\mu(B_d(t, \varepsilon))} \right)^{1/\alpha} d\varepsilon \right). \tag{A.9}$$

where $B_d(t, \varepsilon)$ is the closed ball of center t and radius ε based on the distance d and the infimum is taken over all the probability measure μ on T.

Note that $\gamma_{\alpha,1}(T, d)$ coincides with the functional $\gamma_\alpha(T, d)$ [198] as

$$K(\alpha)^{-1} \gamma_\alpha(T, d) \leq \gamma_{\alpha,1}(T, d) \leq K(\alpha) \gamma_\alpha(T, d), \tag{A.10}$$

where $K(\alpha)$ is a constant depending on α only.

As a result, it is enough to show that $\gamma_{1,1}(T, d_1) \leq \gamma_{2,1}^2(T, d_2)$.

Note that since for any vector x, we have $\|x\|_\infty \leq \|x\|_2$, therefore, for any probability measure μ and t, we have $\mu(B_{d_1}(t, \varepsilon)) \geq \mu(B_{d_2}(t, \varepsilon))$. As a result,

$$\int_0^\infty \left(\log \frac{1}{\mu(B_{d_1}(t, \varepsilon))} \right) d\varepsilon \leq \int_0^\infty \left(\log \frac{1}{\mu(B_{d_2}(t, \varepsilon))} \right) d\varepsilon \tag{A.11}$$

$$\leq \left(\int_0^\infty \left(\log \frac{1}{\mu(B_{d_2}(t,\varepsilon))} \right)^{1/2} d\varepsilon \right)^2 .$$

Since (A.12) holds for any μ and t, we have

$$\begin{aligned} \gamma_{1,1}(T,d_1) &= \inf_t \sup \left(\int_0^\infty \left(\log \frac{1}{\mu(B_{d_1}(t,\varepsilon))} \right) d\varepsilon \right) \\ &\leq \inf_t \sup \left(\int_0^\infty \left(\log \frac{1}{\mu(B_{d_2}(t,\varepsilon))} \right)^{1/2} d\varepsilon \right)^2 \\ &= \gamma_{2,1}^2(T,d_2). \end{aligned} \tag{A.12}$$

This completes the proof. ∎

Theorem 20 ***[Theorem 1.2.7] in [199]*** *Consider a set T provided with two distances d_1 and d_2. Consider a process $(X_t)_{t\in T}$ that satisfies $E[X_t] = 0$ and*

$$P\left(|X_s - X_t| \geq u\right) \leq 2\exp\left(-\min\left(\frac{u^2}{d_2(s,t)^2}, \frac{u}{d_1(s,t)} \right) \right). \tag{A.13}$$

Then

$$E[\sup_{t,s\in T} |X_s - X_t| \leq L(\gamma_1(T,d_1) + \gamma_2(T,d_2)), \tag{A.14}$$

where L is a constant.

Theorem 21 ***[Theorem 1.2.9] in [199]*** *Under the conditions of Theorem 20, for all values of $u_1, u_2 > 0$ we have*

$$\begin{aligned} &P\left(|X_s - X_{t0}| \geq L(\gamma_1(T,d_1) + \gamma_2(T,d_2)) + u_1 D_1 + u_2 D_2\right) \\ &\quad \leq L\exp(-\min(u_2^2, u_1)), \end{aligned} \tag{A.15}$$

where $D_j = 2\sum_{n\geq 0} e_n(T,d_j)$. Note that $D_j \leq L\gamma_j(T,d_j)$.

Theorem 22 ***[Theorem 8.2 (Fernique-Talagrand's comparison theorem)] in [209]*** *Let T be an arbitrary set. Consider a Gaussian random process $(G(t))_{t\in T}$ and*

a sub-Gaussian random process $(H(t))_{t\in T}$. *Assume that* $E[G(t)] = E[H(t)] = 0$ *for all* $t \in T$. *Assume also that for some* $M > 0$, *the following increment comparison holds:*

$$\|\|H(s) - H(t)\|\|_{\psi_2} \leq M(E[\|G(s) - G(t)\|_2^2)^{1/2} \qquad \forall s, t \in T. \tag{A.16}$$

Then

$$E[sup_{t\in T}H(t)] \leq CME[sup_{t\in T}G(t)]. \tag{A.17}$$

Theorem 23 (Mendelson, Pajor, Tomczak-Jaegermann [?]) *There exist absolute constants* c_1, c_2, c_3 *for which the following holds. Let* (Ω, μ) *be a probability space, set* F *be a subset of the unit sphere of* $L_2(\mu)$, *i.e.,* $F \subseteq S_{L_2} = \{f : \|\|f\|\|_{L_2} = 1\}$, *and assume that* $\sup_{f\in F} \|\|f\|\|_{\psi_2} \leq \kappa$. *Then, for any* $\theta > 0$ *and* $n \geq 1$ *satisfying*

$$c_1\kappa\gamma_2(F, \|\|\cdot\|\|_{\psi_2}) \leq \theta\sqrt{n}\ , \tag{A.18}$$

with probability at least $1 - \exp(-c_2\theta^2 n/\kappa^4)$,

$$\sup_{f\in F} \left|\frac{1}{n}\sum_{i=1}^{n} f^2(X_i) - E\left[f^2\right]\right| \leq \theta\ . \tag{A.19}$$

Further, if F *is symmetric, then*

$$E\left[\sup_{f\in F} \left|\frac{1}{n}\sum_{i=1}^{n} f^2(X_i) - E\left[f^2\right]\right|\right] \leq c_3 \max\left\{2\kappa\frac{\gamma_2(F, \|\|\cdot\|\|_{\psi_2})}{\sqrt{n}}, \frac{\gamma_2^2(F, \|\|\cdot\|\|_{\psi_2})}{n}\right\} \tag{A.20}$$

A.2 Regularization Parameter

Lemma 24 *Consider two Ising Model with true parameters* θ_1^* *and* θ_2^*. *Let* $d_1, d_2 \gg s$ *where* $\|\theta_1^*\|_0 = d_1$, $\|\theta_2^*\|_0 = d_2$, *and* $\|\delta\theta^*\|_0 = s$. *Assume*

$$\min_{i,j=1\cdots n_1}(|\theta_1^*(i,j)|) \geq \frac{1}{d_1 - 1} - \frac{c_1}{(d_1 - 1)s} \tag{A.21}$$

$$\min_{i,j=1\cdots n_2}(|\theta_2^*(i,j)|) \geq \frac{1}{d_2-1} - \frac{c_2}{(d_2-1)s}, \tag{A.22}$$

where c_1 and c_2 are positive constants. Then the density ratio $r(X=\mathbf{x}|\delta\theta^*)$ is bounded.

Proof: Let $\alpha_1 \leq |\theta_1^*| \leq \beta_1$ and $\alpha_2 \leq |\theta_2^*| \leq \beta_2$. Without loss of generality, assume that $\|\theta_1^*\|_2 = 1$ and $\|\theta_2^*\|_2 = 1$.

So,

$$\beta_1 \leq 1 - (d_1-1)\alpha_1 \tag{A.23}$$

$$\beta_2 \leq 1 - (d_2-1)\alpha_2. \tag{A.24}$$

Based on triangle inequality of norms, we have

$$\|\delta\theta^*\|_\infty = \|\theta_1^* - \theta_2^*\|_\infty \leq \|\theta_1^*\|_\infty + \|\theta_2^*\|_\infty \leq \beta_1 + \beta_2 \leq 2 - (d_1-1)\alpha_1 - (d_2-1)\alpha_2. \tag{A.25}$$

Let $\mathbf{z} = T(\mathbf{x})$, then,

$$|\langle \mathbf{z}, \delta\theta^* \rangle| \leq \|\mathbf{z}\|_\infty \|\delta\theta^*\|_1 \tag{A.26}$$

$$\leq s\|\delta\theta^*\|_\infty \tag{A.27}$$

$$\leq 2s - [(d_1-1)\alpha_1 - (d_2-1)\alpha_2]s \tag{A.28}$$

where the second inequality is the result of $\|\mathbf{z}\|_\infty \leq 1$ since $\mathbf{z}$ comes from an Ising model.

Note that if

$$\alpha_1 \geq \frac{s-c_1}{(d_1-1)s} = \frac{1}{d_1-1} - \frac{c_1}{(d_1-1)s}, \tag{A.29}$$

then

$$s - (d_1-1)\alpha_1 s \leq c_1. \tag{A.30}$$

Similarly, if

$$\alpha_2 \geq \frac{s-c_2}{(d_2-1)s} = \frac{1}{d_2-1} - \frac{c_2}{(d_2-1)s}, \tag{A.31}$$

then

$$s - (d_2 - 1)\alpha_2 s \leq c_2. \tag{A.32}$$

As a result, we have

$$|\langle \mathbf{z}, \delta\theta^* \rangle| \leq c_1 + c_2. \tag{A.33}$$

$$\Rightarrow \exp\{\langle \mathbf{z}, \delta\theta^* \rangle\} \leq \exp c_1 + c_2 \leq c_0. \tag{A.34}$$

For example, if $c_1 = c_2 = 1$, then $c_0 = 10$.

Therefore,

$$r(X = \mathbf{x}|\delta\theta) = \frac{\exp\{\langle \mathbf{z}, \delta\theta \rangle)\}}{Z(\delta\theta^*)} \leq \frac{c_0}{Z(\delta\theta^*)}. \tag{A.35}$$

This completes the proof. ■

Assumption 1(Smooth Density Ratio Model Assumption) For any vector $\mathbf{u}$ such that $\|\mathbf{u}\|_2 \leq \|\delta\theta^*\|_2$ and every $t \in R$, the following inequality holds:

$$E_{X \sim p(X|\theta_2)}[\exp\{tr(X|\delta\theta^* + \mathbf{u}) - 1\}] \leq \exp\{t^2\}. \tag{A.36}$$

Lemma 25 *For any constant $\tau \leq 1$, define random event M_τ as follows,*

$$M_\tau = \{\Psi(\delta\theta^* + \mathbf{u}) - \Psi(\delta\theta^*) - \left[\hat{\Psi}(\delta\theta^* + \mathbf{u}) - \hat{\Psi}(\delta\theta^*)\right] \leq \tau\}. \tag{A.37}$$

Then, for any vector $\mathbf{u}$ such that $\|\mathbf{u}\|_2 \leq \|\delta\theta^*\|_2$, under Assumption 1, we have

$$P(M_\tau^c) = p\left(\Psi(\delta\theta^* + \mathbf{u}) - \Psi(\delta\theta^*) - \left[\hat{\Psi}(\delta\theta^* + \mathbf{u}) - \hat{\Psi}(\delta\theta^*)\right] > \tau\right) \leq 4e^{-\frac{n_2}{5}\tau^2}. \tag{A.38}$$

Proof: Recall that

$$r(X = \mathbf{x}|\delta\theta^*) = \frac{\exp\{\langle T(\mathbf{x}), \delta\theta^* \rangle\}}{Z(\delta\theta^*)}$$

$$\Rightarrow \qquad \hat{Z}(\delta\theta^*) = \frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle T(\mathbf{x}_i^2), \delta\theta^*\rangle\} = \frac{1}{n_2}\sum_{i=1}^{n_2} r(X = \mathbf{x}_i^2|\delta\theta)Z(\delta\theta^*)$$

$$\Rightarrow \qquad \frac{\hat{Z}(\delta\theta^*)}{Z(\delta\theta^*)} = \frac{1}{n_2}\sum_{i=1}^{n_2} r(X = \mathbf{x}_i^2|\delta\theta^*) \tag{A.39}$$

Note that $Z(\delta\theta) = E_{X\sim p(X|\theta_2)}[\exp\{\langle T(\mathbf{x}), \delta\theta^*\rangle\}]$, therefore,

$$E_{X\sim p(X|\theta_2)}[r(X|\delta\theta^*)] = 1. \tag{A.40}$$

Under the Assumption 1, we have

$$p\left(\left|r(X = \mathbf{x}_i^2|\delta\theta^*) - 1\right| > \epsilon\right) \leq c_1 e^{-\epsilon^2}. \tag{A.41}$$

Applying Hoeffding inequality, we have

$$p\left(\left|\frac{1}{n_2}\sum_{i=1}^{n_2} r(X = \mathbf{x}_i^2|\delta\theta^*) - 1\right| \geq \epsilon\right) \leq 2e^{-\epsilon^2} \tag{A.42}$$

$$\Rightarrow \qquad p\left(\left|\frac{\hat{Z}(\delta\theta^*)}{Z(\delta\theta^*)} - 1\right| \geq \epsilon\right) \leq 2e^{-n_2\epsilon^2}. \tag{A.43}$$

Taking the logarithm from both side, and using one side bound, we have

$$p\left(\log\frac{\hat{Z}(\delta\theta^*)}{Z(\delta\theta^*)} \geq \log(\epsilon + 1)\right) \leq e^{-n_2\epsilon^2} \tag{A.44}$$

$$\Rightarrow p\left(\hat{\Psi}(\delta\theta^*) - \Psi(\delta\theta^*) \geq \log(\epsilon + 1)\right) \leq e^{-n_2\epsilon^2}. \tag{A.45}$$

Similarly, we have

$$p\left(\Psi(\delta\theta^* + \mathbf{u}) - \hat{\Psi}(\delta\theta^* + \mathbf{u}) \geq -\log(1 - \epsilon)\right) \leq e^{-n_2\epsilon^2}. \tag{A.46}$$

Applying the union bound, we have

$$p\left(\Psi(\delta\theta^* + \mathbf{u}) - \Psi(\delta\theta^*) - \left[\hat{\Psi}(\delta\theta^* + \mathbf{u}) - \hat{\Psi}(\delta\theta^*)\right] \geq \log\frac{1+\epsilon}{1-\epsilon}\right) \leq 4e^{-n_2\epsilon^2}. \tag{A.47}$$

Setting $\tau = \log \frac{1+\epsilon}{1-\epsilon}$, we have

$$p\left(\Psi(\delta\theta^* + \mathbf{u}) - \Psi(\delta\theta^*) - \left[\hat{\Psi}(\delta\theta^* + \mathbf{u}) - \hat{\Psi}(\delta\theta^*)\right] \geq \tau\right) \leq 4e^{-n_2\left(\frac{e^\tau+1}{e^\tau-1}\right)^2} \leq 4e^{-\frac{n_2}{5}\tau^2}, \tag{A.48}$$

where the last inequality is obtained by using the fact that for any $\tau \leq 1$

$$\left(\frac{e^\tau+1}{e^\tau-1}\right)^2 > \frac{\tau^2}{5}. \tag{A.49}$$

This completes the proof. ∎

Lemma 26 *Define random event $M_{\tilde{t}}$ as follows,*

$$M_{\tilde{t}} = \{\Psi(\delta\theta^* + t\mathbf{u}) - \Psi(\delta\theta^*) - \left[\hat{\Psi}(\delta\theta^* + t\mathbf{u}) - \hat{\Psi}(\delta\theta^*)\right] \leq \tilde{t}\}, \tag{A.50}$$

where $\tilde{t} = \sqrt{5\eta_1}t + \frac{\sqrt{5}}{2}$. Let $Z = T(X^1)$ and $\mathbf{z} = T(\mathbf{x}^1)$. Then,

$$P\left(\left|\left\langle \mathbf{z} - \nabla\hat{\Psi}(\theta^*), \mathbf{u}\right\rangle\right| \geq \epsilon \middle| M_{\tilde{t}}\right) \leq c\exp\{\frac{-\epsilon^2}{4\eta_0\|\mathbf{u}\|_2^2}\}, \tag{A.51}$$

where $\frac{1}{2}\lambda_{\max}\left(\nabla^2\hat{\Psi}(\theta^* + \tilde{\mathbf{u}})\right) \leq \eta_0$ and c is a positive constant.

Proof: First, note that $p(X = \mathbf{x}|\theta_1^*) = p(X = \mathbf{x}|\theta_2^*)r(X = \mathbf{x}|\delta\theta^*)$. Therefore,

$$\begin{aligned} E_{X\sim p(X|\theta_1^*)}\left[e^{\langle Z, t\mathbf{u}\rangle}\right] &= \sum_{\mathbf{x}\in\mathcal{X}} e^{\langle \mathbf{z}, t\mathbf{u}\rangle} p(\mathbf{x}|\theta_1^*) \\ &= \sum_{\mathbf{x}\in\mathcal{X}} e^{\langle \mathbf{z}, t\mathbf{u}\rangle} p(\mathbf{x}|\theta_2^*) r(\mathbf{x}|\delta\theta^*) \\ &= e^{-\Psi(\delta\theta^*)} \sum_{\mathbf{x}\in\mathcal{X}} e^{\langle \mathbf{z}, t\mathbf{u}+\delta\theta^*\rangle} p(\mathbf{x}|\theta_2^*) \\ &= e^{\Psi(\delta\theta^*+t\mathbf{u})-\Psi(\delta\theta^*)}, \end{aligned} \tag{A.52}$$

since $r(\mathbf{x}|\delta\theta^*) = \exp\{\mathbf{x}, \delta\theta^*\rangle - \Psi(\delta\theta^*)\}$, and $\Psi(\delta\theta^*) = \log\sum_{\mathbf{x}\in\mathcal{X}} e^{\langle \mathbf{x}, \delta\theta^*\rangle} p(\mathbf{x}|\theta_2^*)$.

Also, using the Taylor expansion, we have

$$\hat{\Psi}(\delta\theta^* + t\mathbf{u}) - \hat{\Psi}(\delta\theta^*) - \left\langle \nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle = \frac{1}{2}t\mathbf{u}^T\nabla^2\hat{\Psi}(\delta\theta^* + \tilde{\mathbf{u}})t\mathbf{u}$$

$$\leq \frac{1}{2}t^2\|\mathbf{u}\|_2^2\lambda_{\max}\left(\nabla^2\hat{\Psi}(\delta\theta^*+\tilde{\mathbf{u}})\right) \leq t^2\eta_0\|\mathbf{u}\|_2^2 = t^2\eta_1, \tag{A.53}$$

where $\frac{1}{2}\lambda_{\max}\left(\nabla^2\hat{\Psi}(\delta\theta^*+\tilde{\mathbf{u}})\right) \leq \eta_0$ and $\eta_1 = \eta_0\|\mathbf{u}\|_2^2$.

Then, given the event $M_{\tilde{t}}$, the moment generating function of $\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle$ can be bounded as,

$$\begin{aligned}
&E_{X\sim p(X|\theta_2^*)}\left[E_{X\sim p(X|\theta_1^*)}\left[e^{\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle}\right]\Big|M_{\tilde{t}}\right] \\
&= E_{X\sim p(X|\theta_2^*)}\left[e^{\left\langle -\nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle}\ E_{X\sim p(X|\theta_1^*)}\left[e^{\langle Z, t\mathbf{u}\rangle}\right]\Big|M_{\tilde{t}}\right] \\
&\overset{(A.52)}{=} E_{X\sim p(X|\theta_2^*)}\left[e^{\Psi(\delta\theta^*+t\mathbf{u})-\Psi(\delta\theta^*)-\left\langle\nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle}\Big|M_{\tilde{t}}\right] \\
&\overset{(A.50)}{\leq} E_{X\sim p(X|\theta_2^*)}\left[e^{\hat{\Psi}(\delta\theta^*+t\mathbf{u})-\hat{\Psi}(\delta\theta^*)-\left\langle\nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle+\sqrt{\eta_1}t}\Big|M_{\tilde{t}}\right] \\
&\overset{(A.53)}{\leq} E_{X\sim p(X|\theta_2^*)}\left[e^{t^2\eta_1+\sqrt{5\eta_1}t+\frac{1}{2}}\Big|M_{\tilde{t}}\right] = e^{t^2\eta_1+\sqrt{5\eta_1}t+\frac{1}{2}}.
\end{aligned} \tag{A.54}$$

As a result, using the Chernoff bound, for any $t>0$, we have

$$\begin{aligned}
P\left(\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle \geq \epsilon\Big|M_{\tilde{t}}\right) &\leq e^{-t\epsilon}E_{X\sim p(X|\theta_2^*)}\left[E_{X\sim p(X|\theta_1^*)}\left[e^{\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), t\mathbf{u}\right\rangle}\right]\Big|M_{\tilde{t}}\right] \\
&\overset{(A.54)}{\leq} \exp\{t^2\eta_1+\sqrt{5\eta_1}t+\frac{1}{2}-t\epsilon\} \\
&\overset{(a)}{\leq} \exp\{-\frac{(\epsilon-\sqrt{5\eta_1})^2}{4\eta_1}+\frac{1}{2}\} \\
&\leq c\exp\{-\frac{\epsilon^2}{4\eta_0\|\mathbf{u}\|_2^2}\},
\end{aligned} \tag{A.55}$$

where the inequality (a) is obtained by setting $t=\frac{\epsilon-\sqrt{\eta_1}}{2\eta_1}$ to minimize it with respect to t, and the last inequality obtained by setting $c=\exp\{\frac{5\sqrt{5}}{2\sqrt{\eta_1}}-\frac{3}{4}\}$ and using the fact that

$$\exp\{-\frac{(\epsilon-\sqrt{5\eta_1})^2}{4\eta_1}+\frac{1}{2}\} \leq c\exp\{-\frac{\epsilon^2}{4\eta_1}\}. \tag{A.56}$$

Similarly, we have

$$P\left(\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle \leq -\epsilon\Big|M_{\tilde{t}}\right) \leq e^{-t\epsilon}E_{X\sim p(X|\theta_2^*)}\left[E_{X\sim p(X|\theta_1^*)}\left[e^{\left\langle Z-\nabla\hat{\Psi}(\delta\theta^*), -t\mathbf{u}\right\rangle}\right]\Big|M_{\tilde{t}}\right]$$

$$\leq c \exp\{-\frac{\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\}. \tag{A.57}$$

This completes the proof. ∎

Lemma 27 *Under the smooth density ratio assumption, we have*

$$P\left(\left|\, \langle \nabla \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u} \rangle \,\right| \geq \epsilon\right) \leq c_1 \exp\{-\frac{\min(n_1, n_2)\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\}, \tag{A.58}$$

where c_1 is a positive constant.

Proof: Let $\tilde{t} = \sqrt{5\eta_1}t + \frac{\sqrt{5}}{2}$ and $t = \frac{\epsilon - \sqrt{\eta_1}}{2\eta_1}$. Using the result of lemma 26 we have

$$\begin{aligned} P\left(\left\langle T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon \middle| M_{\tilde{t}}\right) &= P\left(\left\langle T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon \middle| M_{\tilde{t}}\right) \\ &\leq c \exp\{-\frac{\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\}. \end{aligned} \tag{A.59}$$

Applying Hoeffding inequality, we have

$$\begin{aligned} P\left(\left|\, \langle \nabla \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u} \rangle \,\right| \geq \epsilon \middle| M_{\tilde{t}}\right) &= P\left(\left\langle \frac{1}{n_1}\sum_{i=1}^{n_1} T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon \middle| M_{\tilde{t}}\right) \\ &\leq c \exp\{-\frac{n_1\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\}. \end{aligned} \tag{A.60}$$

Moreover, we can obtain,

$$\begin{aligned} P\left(\langle \nabla \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u} \rangle \leq -\epsilon\right) &= P\left(\left\langle \frac{1}{n_1}\sum_{i=1}^{n_1} T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon\right) \\ &\leq P\left(\left\langle \frac{1}{n_1}\sum_{i=1}^{n_1} T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon \middle| M_{\tilde{t}}\right) P(M_{\tilde{t}}) \\ &\quad + P\left(\left\langle \frac{1}{n_1}\sum_{i=1}^{n_1} T(\mathbf{x}_i^1) - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u} \right\rangle \geq \epsilon \middle| M_{\tilde{t}}^c\right) P(M_{\tilde{t}}^c) \\ &\leq c \exp\{\frac{-n_1\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\} + 4\exp\{\frac{-n_2\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\} \\ &\leq c_1 \exp\{-\frac{\min(n_1, n_2)\epsilon^2}{4\eta_0 \|\mathbf{u}\|_2^2}\}, \end{aligned} \tag{A.61}$$

where the last inequality is obtained by using Lemma 25 as follows

$$\begin{aligned} P(M_{\tilde{t}}^c) \leq 4\exp\{-\frac{n_2}{5}\tilde{t}^2\} &= 4\exp\{-\frac{n_2}{5}(\sqrt{5\eta_1}t + \frac{\sqrt{5}}{2})^2\} \\ &= 4\exp\{-\frac{n_2}{5}(\sqrt{5\eta_1}\frac{\epsilon' - \sqrt{\eta_1}}{2\eta_1} + \frac{\sqrt{5}}{2})^2\} \\ &= 4\exp\{-n_2\frac{\epsilon^2}{4\eta_1}\}, \end{aligned}$$

where $\eta_1 = \eta_0\|\mathbf{u}\|_2^2$ and setting $c_1 = \max(4, c)$. Similarly,

$$\begin{aligned} P\left(\langle\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u}\rangle \geq \epsilon\right) &= P\left(\left\langle \frac{1}{n_p}\sum_{i=1}^{n_p} Z_i^p - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle \leq -\epsilon\right) \\ &\leq P\left(\left\langle \frac{1}{n_p}\sum_{i=1}^{n_p} Z_i^p - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle \leq -\epsilon \middle| M_{\tilde{t}}\right) P(M_{t^2}) \\ &\quad + P\left(\left\langle \frac{1}{n_p}\sum_{i=1}^{n_p} Z_i^p - \nabla\hat{\Psi}(\delta\theta^*), \mathbf{u}\right\rangle \leq -\epsilon \middle| M_{\tilde{t}}^c\right) P(M_{\tilde{t}}^c) \\ &\leq c_1\exp\{-\frac{\min(n_1, n_2)\epsilon^2}{4\eta_0\|\mathbf{u}\|_2^2}\} \end{aligned} \tag{A.62}$$

This completes the proof. ∎

Theorem 2 *Define* $\Omega_R = \{u : R(u) \leq 1\}$. *Let* $\phi(R) = \sup_{\mathbf{u}} \frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}$. *Assume that for any* $\mathbf{u}$ *that* $\|\mathbf{u}\| \leq \|\theta^*\|$

$$\frac{1}{2}\lambda_{\max}\left(\nabla^2\mathcal{L}(\delta\theta^* + \mathbf{u})\right) \leq \eta_0, \tag{A.63}$$

where $\lambda_{\max}(.)$ *is the maximum eigenvalue. Then under the smooth density ratio assumption, we have*

$$E\left[R^*(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}))\right] \leq \frac{2\sqrt{\eta_0}}{\sqrt{\min(n_1, n_2)}}(c_1 w\left(\Omega_R\right) + \phi(R)). \tag{A.64}$$

and with probability at least $1 - c_2 e^{-\epsilon^2}$

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right) \leq \frac{1}{\sqrt{\min(n_1, n_2)}}\left(c_2(1+\epsilon)w(\Omega_R) + \tau_1\right). \tag{A.65}$$

where c_1 and c_2 are positive constants, $\tau_1 = 2\sqrt{\eta_0}\phi(R)$, and $w(\Omega_R)$ is the Gaussian width of set Ω_R.

Proof: Define $\boldsymbol{\mu} = E[\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})]$. Using the triangle inequality, we have

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right) \leq R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \boldsymbol{\mu}\right) + R^*\left(\boldsymbol{\mu}\right). \tag{A.66}$$

We upper bound two terms as follows. First, consider the first term.

Using the definition of dual norm, we have

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \boldsymbol{\mu}\right) = \sup_{R(\mathbf{u})\leq 1} \langle \nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \boldsymbol{\mu}, \mathbf{u}\rangle. \tag{A.67}$$

Define stochastic process $H(\mathbf{s}) = \langle \nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \boldsymbol{\mu}, \mathbf{s}\rangle$ where $E[H(\mathbf{s})] = 0$. Then, from Lemma 27, we have

$$P\left(H(\mathbf{s}) - H(\mathbf{t}) \geq \epsilon\right) = P\left(\langle \nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \boldsymbol{\mu}, \mathbf{s}-\mathbf{t}\rangle \geq \epsilon\right) \tag{A.68}$$

$$\leq c_1 \exp\{-\frac{\min(n_1, n_2)\epsilon^2}{4\eta_0\|\mathbf{s}-\mathbf{t}\|_2^2}\}. \tag{A.69}$$

Consider the Gaussian process $G(\mathbf{u}) = \langle \mathbf{u}, g\rangle$, indexed by the same set, i.e., $\mathbf{u} \in \Omega_R$, where $g \sim N(0, \mathbb{I}_{d\times d})$ is standard Gaussian vector. Now from definition sub-Gaussian random variables, we have

$$\|\!|H(\mathbf{s}) - H(\mathbf{t})|\!\|_{\psi_2} \leq \frac{2\sqrt{\eta_0}\|\mathbf{s}-\mathbf{t}\|_2}{\sqrt{\min(n_1, n_2)}} = KE_g[\|G(\mathbf{s}) - G(\mathbf{t})\|_2^2]^{1/2}, \tag{A.70}$$

where $E_g[\|G(\mathbf{s}) - G(\mathbf{t})\|_2^2]^{1/2} = E_g[\|\langle \mathbf{s}-\mathbf{t}, g\rangle\|_2^2]^{1/2} = \|\mathbf{s}-\mathbf{t}\|_2$, and $K = \frac{2\sqrt{\eta_0}}{\sqrt{\min(n_1,n_2)}}$.

Next, by applying the Fernique-Talagrand's comparison theorem 22, we have

$$E[\sup_{\mathbf{u}\in\Omega_R} H(\mathbf{u})] = E\left[\sup_{\mathbf{u}\in\Omega_R} \langle \nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u}\rangle\right]$$

$$\leq c_1 K E[\sup_{u\in\Omega_R} G(u)] = 2c_1\sqrt{\eta_0}\frac{w(\Omega_R)}{\sqrt{\min(n_1, n_2)}}, \tag{A.71}$$

where c_1 is a constant. Thus,

$$E\left[R^*\left(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})-\boldsymbol{\mu}\right)\right]\leq c_1\frac{w(\Omega_R)}{\sqrt{\min(n_1,n_2)}}. \tag{A.72}$$

To get the concentration bound, we use the direct application of Theorem 2.2.27 in [200] and we have

$$P\left(\sup_{\mathbf{s},\mathbf{t}\in\Omega_R}|H(\mathbf{s})-H(\mathbf{t})|\leq c_2(1+\epsilon)\frac{w(\Omega_R)}{\sqrt{\min(n_1,n_2)}}\right)\geq 1-c_2\exp\left(-\epsilon^2\right). \tag{A.73}$$

Thus, with probability at least $1-c_2\exp\left(-\epsilon^2\right)$,

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})-\boldsymbol{\mu}\right)\leq c_2(1+\epsilon)\frac{w(\Omega_R)}{\sqrt{\min(n_1,n_2)}}. \tag{A.74}$$

Next, we consider the second term. First note that $\left|\left|\left|\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right|\right|\right|_{\Psi_2}\leq\frac{2\sqrt{\eta_0}\|\mathbf{u}\|_2}{\sqrt{\min(n_1,n_2)}}$. Using sub-Gaussian variables property, we have

$$E[\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})]\leq\left|\left|\left|\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right|\right|\right|_{\Psi_2}\leq\frac{2\sqrt{\eta_0}\|\mathbf{u}\|_2}{\sqrt{\min(n_1,n_2)}} \tag{A.75}$$

Using the definition of the dual norm, we have

$$R^*(\boldsymbol{\mu})=R^*\left(E\left[\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2})\right]\right)=\sup_{\mathbf{u}\in\Omega_R}E\left[\langle\nabla\mathcal{L}(\delta\theta^*;\mathfrak{X}_1^{n_1},\mathfrak{X}_2^{n_2}),\mathbf{u}\rangle\right] \tag{A.76}$$

$$\leq\frac{2\sqrt{\eta_0}}{\sqrt{\min(n_1,n_2)}}\sup_{\mathbf{u}}\frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}=\frac{2\sqrt{\eta_0}}{\sqrt{\min(n_1,n_2)}}\Phi(R), \tag{A.77}$$

where $\Phi(R)=\sup_{\mathbf{u}}\frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}$.

Also, we have

$$E\left[R^*(\boldsymbol{\mu})\right]=\leq\frac{2\sqrt{\eta_0}}{\sqrt{\min(n_1,n_2)}}\Phi(R), \tag{A.78}$$

where $\Phi(R)=\sup_{\mathbf{u}}\frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}$.

This completes the proof. ∎

A.3 RSC condition

Let $r_i = r(X = \mathbf{x_i^2}|\delta\theta^*)$ and $\bar{\varepsilon}$ denote the probability that r_i exceeds some constant η_0: $\bar{\varepsilon} = p(r_i > \eta_0) \geq 1 - e^{-\frac{\eta_0^2}{2}}$.

Theorem 5 *Let $X \in \mathbb{R}^{n\times p}$ be a design matrix with independent isotropic sub-Gaussian rows with $|||X_i|||_{\Psi_2} \leq \kappa$. Then, for any set $A \subseteq S^{p-1}$, for suitable constants η, c_1, $c_2 > 0$ with probability at least $1 - \exp(-\eta w^2(A))$, we have*

$$\inf_{u\in A} \partial\mathcal{L}(\theta^*; u, X) \geq c_1\underline{\rho}^2 \left(1 - c_2\kappa_1^2 \frac{w(A)}{\sqrt{n_2}}\right) - \tau \tag{A.79}$$

where $\kappa_1 = \frac{\kappa}{\bar{\varepsilon}}$, $\underline{\rho}^2 = \inf_{\mathbf{u}\in A} \rho_{\mathbf{u}}^2$ with $\rho_{\mathbf{u}}^2 = E\left[\left\langle \mathbf{u}, T(X_i^2)\right\rangle^2 \mathbb{I}(r_i > \eta_0)\right]$, and τ is smaller than the first term in right hand side. Thus, for $n_2 \geq c_2 w^2(A)$, with probability at least $1 - \exp(-\eta w^2(A))$, we have $\inf_{u\in A} \partial\mathcal{L}(\theta^*; u, X) > 0$.

Proof: Define $Z = T(X)$ and $\mathbf{z}_i = T(\mathbf{x}_i^2)$. Then,

$$\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) = \frac{-1}{n_1}\sum_{i=1}^{n_1}\langle T(\mathbf{x}_i^1), \delta\theta\rangle + \log\frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle T(\mathbf{x}_i^2), \delta\theta\rangle\} \tag{A.80}$$

$$= \frac{-1}{n_1}\sum_{i=1}^{n_1}\langle \mathbf{z_i}, \delta\theta\rangle + \log\frac{1}{n_2}\sum_{i=1}^{n_2}\exp\{\langle \mathbf{z}_i, \delta\theta\rangle\}. \tag{A.81}$$

Through the analysis, we consider that Z is centered random variable without loss of generality, since if it is not, the $E[Z]$ will show up as a constant.

Recall, RSC condition definition as

$$\delta\mathcal{L}(\delta\theta^*, \mathbf{u}) := \mathcal{L}(\delta\theta^* + \mathbf{u}; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \langle\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u}\rangle \geq \kappa\|\mathbf{u}\|_2^2 \tag{A.82}$$

Simplifying the expression and applying mean value theorem twice on the left side of RSC condition (3.26), for $\forall\gamma_i \in [0, 1]$, we have

$$\delta\mathcal{L}(\delta\theta^*, \mathbf{u}) := \mathcal{L}(\delta\theta^* + \mathbf{u}; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) - \langle\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}), \mathbf{u}\rangle$$

$$\geq \mathbf{u}^T \nabla^2 \mathcal{L}(\delta\tilde{\theta}; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\mathbf{u}, \tag{A.83}$$

where $\delta\tilde{\theta} = \delta\theta^* + \gamma_i \mathbf{u}$. As a result, to show when the RSC condition is satisfied, it is enough to find a lower bound for the right side of the above equation.

Note that

$$\nabla^2 \mathcal{L}(\delta\tilde{\theta}; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) = \nabla^2 \hat{\Psi}(\delta\tilde{\theta}), \tag{A.84}$$

where

$$\nabla^2 \hat{\Psi}(\delta\tilde{\theta}) = \sum_{i=1}^{n_2} \sigma_i \mathbf{z}_i^T \mathbf{z}_i - \left(\sum_{j=1}^{n_2} \sigma_j \mathbf{z}_j \right)^T \left(\sum_{j=1}^{n_2} \sigma_j \mathbf{z}_j \right), \tag{A.85}$$

and

$$\sigma_i = \exp\{\langle \mathbf{z}_i, \delta\tilde{\theta} \rangle - \hat{\Psi}(\delta\tilde{\theta})\} = \frac{\exp \langle \mathbf{z}_i, \delta\tilde{\theta} \rangle}{\sum_{j=1}^{n_2} \exp \langle \mathbf{z}_j, \delta\tilde{\theta} \rangle}. \tag{A.86}$$

Putting (A.85) back in (A.83), we have

$$\delta\mathcal{L}(\delta\theta^*, \mathbf{u}) \geq \underbrace{\sum_{i=1}^{n_2} \sigma_i \langle \mathbf{u}, \mathbf{z}_i \rangle^2}_{A} - \underbrace{\left\langle \mathbf{u}, \sum_{j=1}^{n_2} \sigma_j \mathbf{z}_j \right\rangle^2}_{B}. \tag{A.87}$$

To show the RSC condition, we need to show that (A.87) is strictly positive. First, we obtain the sample complexity so that A is far away from zero, then we show that A is strictly greater than B. This is enough to obtain the sample complexity so that the RSC condition is satisfied.

i. Lower bound on A: Here, we explain how to get a lower bound on $\inf_{\mathbf{u} \in A} \sum_{i=1}^{n_2} \sigma_i \langle \mathbf{u}, \mathbf{z}_i \rangle^2$. Let $r_i = r(X = \mathbf{x_i^2} | \delta\theta^*)$, and $s_r = \sum_{j=1}^{n_2} r_j$, then $\sigma_i = \frac{r_i}{s_r}$. Then, we have

$$\sum_{i=1}^{n_2} \sigma_i \langle \mathbf{u}, \mathbf{z}_i \rangle^2 = \frac{1}{s_r} \sum_{i=1}^{n_2} r_i \langle \mathbf{u}, \mathbf{z}_i \rangle^2 \ . \tag{A.88}$$

Then, we have

$$p\left(\inf_{\mathbf{u}\in A}\frac{1}{s_r}\sum_{i=1}^{n_2} r_i\langle \mathbf{u},\mathbf{z}_i\rangle^2 < \frac{\eta_0}{\eta_1}\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\right) \le p(\frac{1}{s_r} < \frac{1}{\eta_1 n_2}) + p\left(\inf_{\mathbf{u}\in A}\sum_{i=1}^{n_2} r_i\langle \mathbf{u},\mathbf{z}_i\rangle^2 < \eta_0 n_2\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\right). \quad \text{(A.89)}$$

First, we give a bound for the first term. Note that $E_{X\sim p(X|\theta_2)}[r(X|\delta\theta^*)] = 1$. From the smooth density ratio model assumption, we have

$$p(|r_i - 1| > t) \le 2e^{-\frac{t^2}{2}}. \quad \text{(A.90)}$$

Applying Hoeffding inequality in (A.94), we have

$$p(|\frac{1}{n_2}s_r - 1| \ge t) = p(|\frac{1}{n_2}\sum_{j=1}^{n_2} r_j - 1| \ge t) \le 2e^{-\frac{n_2 t^2}{2}}, \quad \text{(A.91)}$$

$$\Rightarrow \quad p(s_r \ge \eta_1 n_2) \le e^{-\frac{n_2(\eta_1-1)^2}{2}}, \quad \text{(A.92)}$$

$$\Rightarrow \quad p(\frac{1}{s_r} \le \frac{1}{\eta_1 n_2}) = p(s_r \ge \eta_1 n_2) \le e^{-\frac{n_2(\eta_1-1)^2}{2}}, \quad \text{(A.93)}$$

where $\eta_1 = t + 1$.

Next, we focus on bounding the second term in (A.89). Recall that,

$$p(|r_i - 1| > t) \le 2e^{-\frac{t^2}{2}}, \quad \text{(A.94)}$$

$$\Rightarrow \quad \bar{\varepsilon}_1 = p(r_i \ge \eta_0) \ge 1 - e^{-\frac{(1-\eta_0)^2}{2}}, \quad \text{(A.95)}$$

where the last inequality holds for any $\eta_0 = 1 - t$.

For any fixed η_0, let $\bar{W}_i = \bar{W}_i^u = \langle \mathbf{u},\mathbf{z}_i\rangle \mathbb{I}(r_i > \eta_0)$. Then, the probability distribution over $\bar{W}_i$ can be written as:[1]

$$p(\bar{W}_i = w) = \frac{p(\langle \mathbf{u},\mathbf{z}_i\rangle = w)\mathbb{I}(r_i > \eta_0)}{p(r_i > \eta_0)} \le \frac{1}{\bar{\varepsilon}_1}p(\langle \mathbf{u},\mathbf{z}_i\rangle = w)\ . \quad \text{(A.96)}$$

[1]With abuse of notation, we treat the distribution over $\bar{W}_i$ as discrete for ease of notation. A similar argument applies for the true continuous distribution, but more notation is needed.

As a result, $\left|\left|\left|\bar{W}_i\right|\right|\right|_{\psi_2} \leq \frac{\kappa}{\bar{\varepsilon}_1} = \kappa_1$. Thus, $\bar{W}_i = \bar{W}_i^u$ is a sub-Gaussian random variable for any $\mathbf{u} \in A$. Let $\rho_{\mathbf{u}}^2 = E[(\bar{W}_i{}^u)^2] > 0$. For convenience of notation, let Z_0 be i.i.d. as the rows $\mathbf{z}_i, i = 1, \ldots, n$. Let $A \subseteq S^{p-1}$. Consider the following class of functions:

$$F = \{f_{\mathbf{u}}, \mathbf{u} \in A : f_{\mathbf{u}}(.) = \frac{1}{\rho_{\mathbf{u}}} \langle \cdot, \mathbf{u} \rangle \mathbb{I}(r(.|\delta\theta^*) \geq \eta_0) : \mathbf{u} \in A\}. \tag{A.97}$$

Then for any $f_{\mathbf{u}} \in F$, $f_{\mathbf{u}}(Z_0) = \frac{1}{\rho_{\mathbf{u}}} \langle Z_0, \mathbf{u} \rangle \mathbb{I}(r_i \geq \eta_0)$ and, by construction, F is a subset of the unit sphere, since for $f_{\mathbf{u}} \in F$

$$\|f_{\mathbf{u}}\|_{L_2}^2 = \frac{1}{\rho_{\mathbf{u}}^2} E[\langle Z_0, \mathbf{u} \rangle^2 \mathbb{I}(r_i \geq \eta_0)] = 1. \tag{A.98}$$

Further, $\sup_{f_{\mathbf{u}} \in F} \left|\left|\left| f_{\mathbf{u}} \right|\right|\right|_{\psi_2} \leq \kappa_1/2$.

Next, we show that for the current setting, the γ_2-functional can be upper bounded by $w(A)$, the Gaussian width of A. Since the process is sub-Gaussian with φ_2-norm bounded by κ_1, we have

$$\gamma_2(F \cap S_{L_2}, \left|\left|\left| \cdot \right|\right|\right|_{\psi_2}) \leq \kappa_1 \gamma_2(F \cap S_{L_2}, \left|\left|\left| \cdot \right|\right|\right|_{L_2}) \leq \kappa_1 c_4 w(A) \ , \tag{A.99}$$

where the last inequality follows from generic chaining, in particular [199, Theorem 2.1.1], for an absolute constant $c_4 > 0$.

In the context of Theorem 23, we choose

$$\theta = c_1 c_4 \kappa_1^2 \frac{w(A)}{\sqrt{n}} \geq c_1 \kappa_1 \frac{\gamma_2(F \cap S_{L_2}, \left|\left|\left| \cdot \right|\right|\right|_{\varphi_2})}{\sqrt{n}}, \tag{A.100}$$

so that the condition on θ is satisfied. With this choice of θ, we have

$$\theta^2 n / \kappa_1^4 = c_1^2 c_4^2 w^2(A) \ . \tag{A.101}$$

Then, from Theorem 23, it follows that with probability at least $1 - \exp(-\eta w^2(A))$, we have

$$\sup_{\mathbf{u} \in A} \left| \frac{1}{\rho_{\mathbf{u}} n_2} \sum_{i=1}^{n_2} \frac{1}{\rho_{\mathbf{u}}} \langle \mathbf{z}_i, \mathbf{u} \rangle^2 \mathbb{I}(r_i \geq \eta_0) - 1 \right| \leq c \kappa_1^2 \frac{w(A)}{\sqrt{n_2}}, \tag{A.102}$$

where $\eta = c_2 c_1^2 c_4^2$ and $c = c_1 c_2$ are absolute constants. Thus, with probability at least

$1-\exp(-\eta w^2(A))$,

$$\inf_{\mathbf{u}\in A}\frac{1}{n_2}\sum_{i=1}^{n_2}\langle \mathbf{z}_i,\mathbf{u}\rangle^2\mathbb{I}(r_i\geq\eta_0)\geq\inf_{\mathbf{u}\in A}\ \rho_{\mathbf{u}}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right), \tag{A.103}$$

$$\Rightarrow\quad \inf_{\mathbf{u}\in A}\sum_{i=1}^{n_2}\langle \mathbf{z}_i,\mathbf{u}\rangle^2\mathbb{I}(r_i\geq\eta_0)\geq n_2\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right), \tag{A.104}$$

where $\underline{\rho}^2=\inf_{\mathbf{u}\in A}\rho_{\mathbf{u}}^2$. Then, with probability at least $1-\exp(-\eta w^2(A))$, we have

$$\inf_{u\in A}\sum_{i=1}^{n_2}r_i\langle \mathbf{z}_i,\mathbf{u}\rangle^2\geq\inf_{u\in A}\sum_{i=1}^{n_2}r_i\langle \mathbf{z}_i,\mathbf{u}\rangle^2\mathbb{I}(r_i\geq\eta_0) \tag{A.105}$$

$$\geq\inf_{u\in A}\eta_0\sum_{i=1}^{n_2}\langle \mathbf{z}_i,\mathbf{u}\rangle^2\mathbb{I}(r_i\geq\eta_0) \tag{A.106}$$

$$\geq\eta_0 n_2\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\ . \tag{A.107}$$

Thus,

$$p\left(\inf_{u\in A}\sum_{i=1}^{n_2}r_i\langle \mathbf{z}_i,\mathbf{u}\rangle^2\geq\eta_0 n_2\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\right)\geq 1-\exp(-\eta w^2(A)), \tag{A.108}$$

$$\Rightarrow\quad p\left(\inf_{u\in A}\sum_{i=1}^{n_2}r_i\langle \mathbf{z}_i,\mathbf{u}\rangle^2<\eta_0 n_2\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\right)\leq\exp(-\eta w^2(A)). \tag{A.109}$$

Putting (A.93) and (A.109) into (A.89), for any $n_2\geq\frac{2\eta w^2(A)}{(\eta_1-1)^2}$ we have

$$p\left(\inf_{u\in A}\sum_{i=1}^{n_2}\sigma_i\langle \mathbf{z}_i,\mathbf{u}\rangle^2<\frac{\eta_0}{\eta_1}\underline{\rho}^2\left(1-c\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right)\right)\leq\exp(-\frac{n_2(\eta_1-1)^2}{2})+\exp(-\eta w^2(A)) \tag{A.110}$$

$$\leq 2\exp(-\eta w^2(A)), \tag{A.111}$$

ii. A is strictly greater than B: Note that, $0\leq\sigma_i\leq 1$ for all i and $\sum_{i=1}^{n_2}\sigma_i=1$. Define $f(\mathbf{z})=\langle \mathbf{u},\mathbf{z}\rangle^2$, which is a convex function of $\mathbf{z}$. Using Jensen's inequality, we

have

$$f\left(\frac{\sum_{i=1}^{n_2}\sigma_i\mathbf{z}_i}{\sum_{i=1}^{n_2}\sigma_i}\right) \leq \frac{\sum_{i=1}^{n_2}\sigma_i f(\mathbf{z}_i)}{\sum_{i=1}^{n_2}\sigma_i} \tag{A.112}$$

$$\left\langle \mathbf{u}, \frac{\sum_{i=1}^{n_2}\sigma_i\mathbf{z}_i}{\sum_{i=1}^{n_2}\sigma_i}\right\rangle^2 \leq \frac{\sum_{i=1}^{n_2}\sigma_i\langle\mathbf{u},\mathbf{z}_i\rangle^2}{\sum_{i=1}^{n_2}\sigma_i} \tag{A.113}$$

$$\langle\mathbf{u}, \sum_{i=1}^{n_2}\sigma_j\mathbf{z}_i\rangle^2 \leq \sum_{i=1}^{n_2}\sigma_i\langle\mathbf{u},\mathbf{z}_i\rangle^2. \tag{A.114}$$

The equality in (A.114) holds if $\mathbf{z}_1 = \mathbf{z}_2 = \cdots = \mathbf{z}_{n_2}$, or if both sides are zero i.e., $\mathbf{u}$ is in the null space of $\mathbf{z}_i$ for all i. Since $\mathbf{z}_i$ are different with probability 1, then if we show that $\mathbf{u}$ is not in the null space of $\mathbf{z}_i$ for all i, then the inequality (A.114) is strict inequality.

This completes the proof. ∎